Scaled, inexact and adaptive generalized FISTA for strongly convex optimization

Simone Rebegoldi*, Luca Calatroni†

Abstract

We consider a variable metric and inexact version of the FISTA-type algorithm considered in [14, 16] for the minimization of the sum of two (possibly strongly) convex functions. The proposed algorithm is combined with an adaptive (non-monotone) backtracking strategy, which allows for the adjustment of the algorithmic step-size along the iterations in order to improve the convergence speed. We prove a linear convergence result for the function values, which depends on both the strong convexity moduli of the two functions and the upper and lower bounds on the spectrum of the variable metric operators. We validate the proposed algorithm, named Scaled Adaptive GEneralized FISTA (SAGE-FISTA), on exemplar image denoising and deblurring problems where edge-preserving Total Variation (TV) regularization is combined with Kullback-Leibler-type fidelity terms, as it is common in applications where signal-dependent Poisson noise is assumed in the data.

1 Introduction

In recent years, inertial Forward–Backward (FB) methods have become standard tools for solving composite convex optimization problems of the form

$$\min_{x\in\mathcal{H}} F(x) \equiv f(x) + g(x), \tag{1}$$

where $\mathcal{H}$ is an Hilbert space, $f, g : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ are convex functions and f is continuously differentiable on a subset $Y \subseteq \mathcal{H}$. In its standard form, the FB algorithm [17, 19] is defined as the sequential application of a (forward) gradient step on f followed by a (backward) proximal step on g, thus resulting in the following iterative scheme defined for all $k \geq 0$

$$x^{(k+1)} = \mathrm{prox}_{\tau_k g}(x^{(k)} - \tau_k \nabla f(x^{(k)})) = \operatorname*{argmin}_{z\in\mathcal{H}} \; g(z) + \frac{1}{2\tau_k}\|z - x^{(k)} + \tau_k \nabla f(x^{(k)})\|^2, \tag{2}$$

where $\tau_k > 0$ is the algorithmic step-size and $\mathrm{prox}_{\tau_k g}$ denotes the proximal operator associated to the convex function $\tau_k g$ [30]. In order to improve the convergence rate of (2) to a minimum point $x^* \in \operatorname{argmin} F$, one can consider the following inertial variant of the FB scheme [4, 31]:

$$x^{(k+1)} = \mathrm{prox}_{\tau_{k+1} g}(y^{(k+1)} - \tau_{k+1}\nabla f(y^{(k+1)})), \quad \text{where } y^{(k+1)} = x^{(k)} + \beta_{k+1}(x^{(k)} - x^{(k-1)}). \tag{3}$$

Here, $y^{(k+1)}$ denotes the inertial point obtained by extrapolation of the sequence $\{x^{(k)}\}_{k\in\mathbb{N}}$ and β_{k+1} is the inertial parameter. The idea of introducing inertia in (2) traces back to the pioneering

*Dipartimento di Ingegneria Industriale, Università degli studi di Firenze, (simone.rebegoldi@unifi.it).
†Laboratoire I3S, CNRS, UCA, Inria (calatroni@i3s.unice.fr).

work by Polyak [35], where the optimal convergence rate of the Heavy-ball method is discussed for strongly convex and smooth problems, and the one by Nesterov [31], where the author proposes an inertial gradient method of type (3) suited for the case $g \equiv 0$ with optimal convergence rate $\mathcal{O}(1/k^2)$, that is one order higher than the standard $\mathcal{O}(1/k)$ rate available for the method (2). This approach has been extended by Nesterov in [32] to projected gradient descent algorithms and, later, to a general non-smooth convex function g in [4], under the popular name of "Fast Iterative Soft-Thresholding Algorithm" (FISTA). In spite of its optimal convergence rate, the practical acceleration of FISTA is often limited by the selection of the algorithmic step-sizes τ_k, which, typically, are computed either as the inverse of a (pessimistic) estimate of the Lipschitz constant L_f of the gradient of f, or through a monotone Armijo-type backtracking procedure when such an estimate is not available [4, Section 4]. Both choices usually generate a sequence of admissible step-sizes which, in case they are too small, may lead to a severe slowdown in the numerical performance of the algorithm. Possible remedies to this issue consist in either adopting an adaptive (non-monotone) backtracking procedure allowing for the local increasing of the step-size [33,38], or computing the proximal–gradient operator in (3) with respect to the norm induced by a suitably-chosen variable linear operator, which depends on some second-order information of the function f at the current inertial point. Such variable metric strategy has been popularized in recent years and thoroughly studied in a series of works, such as, e.g., [8, 10, 11, 18, 24].

Furthermore, if one or both functions f, g in (1) are strongly convex, then it is possible to incorporate the strong convexity moduli inside the definition of the inertial parameters β_{k+1} in order to obtain linear convergence rates. This property was shown first by Nesterov in [32, Section 2.2.1, Theorem 2.2.3] for smooth optimization problems ($g \equiv 0$) and later extended in [39] for general non-smooth functions g, assuming strong convexity only on the smooth component f. In [16, Algorithm 5], the authors consider a variant of this algorithm for problems where either one or both terms f, g are strongly convex. The same algorithm is denominated Generalised FISTA (GFISTA) in [14] and equipped with an adaptive backtracking strategy allowing for local increasing of the step-size, similarly to the one proposed in [38]. In [22, 23] similar algorithms are considered and thoroughly analyzed by means of generalized estimate sequences arguments. However, the convergence analysis carried out in [14, 16, 22, 23] does not treat explicitly the case where the proximal operator of g is computed inexactly due to the lack of an exact analytical formula. This happens frequently in signal and image processing problems, especially in the presence of sparsity-inducing priors [1]. The practical performance of GFISTA may be in fact severely affected by inexact proximal evaluations, since inertial FB variants are usually more sensitive to computational errors than the standard FB scheme, see, e.g. [39,40]. In this context, the adoption of a variable metric strategy in (3) is expected to compensate for the proximal errors, thus favoring the recovery of the expected linear rate of GFISTA, as observed also for standard FISTA in [11].

Contribution and structure of the paper In this paper, we propose SAGE-FISTA (Scaled Adaptive GEneralized FISTA), an inertial variable metric FB algorithm with inexact proximal evaluations suited for the minimization of composite (possibly strongly) convex functionals. The proposed algorithm computes the FB iterates inexactly with increasing accuracy and with respect to a sequence of suitably-chosen variable metric operators. In order to limit the computational slowdowns due to the use of possibly small step-sizes generated by Armijo-type monotone backtracking strategies, we further consider an adaptive backtracking procedure allowing for local adjustment of such parameter. On the one hand, SAGE-FISTA can be seen as a variable metric and inexact version of the GFISTA algorithm proposed in [14, 16], as the original GFISTA did not include scaling operators nor inexact proximal evaluations. On the other hand, we can interpret SAGE-FISTA as a non-monotone extension of the scaled FISTA-like algorithms devised

in [11, 24], which already featured variable metrics and proximal inexactness, but without the adoption of a non-monotone backtracking procedure for the step-sizes. Thus, our main contribution lies in the simultaneous combination of variable metrics, inexact proximal evaluations and non-monotone backtracking, while taking into account strong convexity. Under the assumption that one or both functions f, g in (1) are strongly convex, we prove the linear convergence of the sequence of function values generated by SAGE-FISTA. Furthermore, we assess the numerical performance of SAGE-FISTA on some image restoration problems, showing that the combination of the scaling technique with the adaptive backtracking strategy significantly reduces the computational times and avoids misspecifications of the algorithmic step-size.

The paper is organized as follows. In Section 2 we set the notation and report some auxiliary technical results. In Section 3 we describe SAGE-FISTA in detail and provide a rigorous study of its convergence properties, showing, in particular, that linear convergence is achieved in strongly convex scenarios. We further report some useful results for the practical definition of both the inexactness parameter sequence and the variable metric operators. Finally, in Section 4, we apply SAGE-FISTA to exemplar image restoration problems such as Poisson image denoising and deblurring.

2 Problem setting and preliminaries

In this section, we fix the notation, formulate the general optimization problem we aim to solve and prove some important preliminary results for the convergence analysis that follows.

2.1 Notations

For a given Hilbert space $\mathcal{H}$, we denote by $\|\cdot\|$ the norm induced by the inner product $\langle\cdot,\cdot\rangle$ on $\mathcal{H}$. For any function $f : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$, we denote the domain of f by $\mathrm{dom}(f) := \{x \in \mathcal{H} : \ f(x) < \infty\}$. If $f : \mathcal{H} \to \mathbb{R}$ is a continuously differentiable function, then by $\mathbb{D}_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y\rangle$ we denote the Bregman distance of f between $x, y \in \mathcal{H}$.
Let $\mathcal{S}(\mathcal{H})$ be the set of linear, bounded and self-adjoint operators from $\mathcal{H}$ to $\mathcal{H}$, and $\mathcal{I} \in \mathcal{S}(\mathcal{H})$ the identity operator on $\mathcal{H}$. The Loewner partial ordering relation on $\mathcal{S}(\mathcal{H})$ is defined as follows:

$$\forall D_1, D_2 \in \mathcal{S}(\mathcal{H}) \quad D_1 \preceq D_2 \Leftrightarrow \langle D_1 x, x\rangle \leq \langle D_2 x, x\rangle \ \forall x \in \mathcal{H}\,.$$

For any η_{inf}, η_{sup} with $0 < \eta_{inf} \leq \eta_{sup}$, we introduce the following sets

$$\mathcal{D}_{\eta_{inf}} := \{D \in \mathcal{S}(\mathcal{H}) : \ \eta_{inf}\mathcal{I} \preceq D\}, \quad \mathcal{D}_{\eta_{inf}}^{\eta_{sup}} := \{D \in \mathcal{S}(\mathcal{H}) : \ \eta_{inf}\mathcal{I} \preceq D \preceq \eta_{sup}\mathcal{I}\}.$$

By definition $\mathcal{D}_{\eta_{inf}}^{\eta_{sup}} \subseteq \mathcal{D}_{\eta_{inf}}$. Furthermore, following [20, Theorem 4.6.11], we have that if $D \in \mathcal{D}_{\eta_{inf}}$, then D is invertible and, if also $D \in \mathcal{D}_{\eta_{inf}}^{\eta_{sup}}$, then we have $D^{-1} \in \mathcal{D}_{1/\eta_{sup}}^{1/\eta_{inf}}$. If $D \in \mathcal{D}_{\eta_{inf}}$, then

$$(x, y) := \langle Dx, y\rangle, \tag{4}$$

defines an inner product on $\mathcal{H}$, and $\|x\|_D = \sqrt{\langle Dx, x\rangle}$ denotes the $D-$norm induced by (4). Hence, if $D \in \mathcal{D}_{\eta_{inf}}^{\eta_{sup}}$, the following inequality holds

$$\eta_{inf}\|x\|^2 \leq \|x\|_D^2 \leq \eta_{sup}\|x\|^2, \quad \forall\ x \in \mathcal{H}. \tag{5}$$

Given $Y \subseteq \mathcal{H}$ nonempty, closed convex set and $D \in \mathcal{D}_{\eta_{inf}}$, we define the projection operator onto Y in the metric induced by D as $P_{Y,D}(x) = \operatorname*{argmin}_{z \in Y} \|z - x\|_D^2$ for all $x \in \mathcal{H}$.

2.2 Problem formulation and preliminary results

We are interested in solving the optimization problem

$$\min_{x\in\mathcal{H}} F(x) \equiv f(x) + g(x)\,, \tag{6}$$

where

- $f : \mathcal{H} \to \mathbb{R}$ is μ_f-strongly convex ($\mu_f \geq 0$) and continuously differentiable with L_f - Lipschitz continuous gradient on a closed convex set $Y \neq \emptyset$ with $\mathrm{dom}(g) \subseteq Y \subseteq \mathrm{dom}(f)$;
- $g : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ is proper, μ_g-strongly convex ($\mu_g \geq 0$), and lower semicontinuous.

Note that by saying that a function f is strongly convex with $\mu_f = 0$, we just mean that f is convex.
In the following, we prove a technical descent inequality (Lemma 2.3) that is crucial for the convergence analysis carried out in Section 3. We start noting that, if f is μ_f-strongly convex, then f is also strongly convex with respect to the norm induced by any operator $D \in \mathcal{D}_{\eta_{inf}}^{\eta_{sup}}$.

Lemma 2.1 *Let $f : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ be a μ_f-strongly convex function on a nonempty convex set $Y \subseteq \mathrm{dom}(f)$. For all $D \in \mathcal{D}_{\eta_{inf}}^{\eta_{sup}}$ and $\mu_{f,D} \leq \frac{\mu_f}{\eta_{sup}}$, we have that f is $\mu_{f,D}-$strongly convex with respect to the $D-$norm, that is*

$$f(y) \geq f(x) + \langle w, y - x\rangle + \frac{\mu_{f,D}}{2}\|y - x\|_D^2, \quad \forall\, y, x \in Y,\ w \in \partial f(x). \tag{7}$$

Proof. In view of the μ_f-strong convexity of f and (5), for all $x, y \in Y$, $w \in \partial f(x)$ we have

$$f(y) \geq f(x) + \langle w, y - x\rangle + \frac{\mu_f}{2\eta_{sup}}\eta_{sup}\|y - x\|^2 \geq f(x) + \langle w, y - x\rangle + \frac{\mu_{f,D}}{2}\|y - x\|_D^2.$$

□

We now report the statement of a scaled version of the descent lemma (see [10] for a proof).

Lemma 2.2 *[10, Lemma 6] Let $f : \mathcal{H} \to \mathbb{R}\cup\{\infty\}$ be a continuously differentiable function with L_f-Lipschitz continuous gradient on $Y \subseteq \mathrm{dom}(f)$ and $D \in \mathcal{D}_{\eta_{inf}}$. If $0 < \tau \leq \frac{\eta_{inf}}{L_f}$, we have*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2\tau}\|x - y\|_D^2 \quad \forall\, x, y \in Y. \tag{8}$$

For $\tau > 0$, $D \in \mathcal{D}_{\eta_{inf}}^{\eta_{sup}}$, we define the proximal operator of g in the metric induced by D as

$$\mathrm{prox}_g^D(x) = \operatorname*{argmin}_{z\in\mathcal{H}} g(z) + \frac{1}{2}\|z - x\|_D^2, \quad \forall\, x \in \mathcal{H}.$$

Furthermore, given $\bar{x} \in Y$, we introduce the function $h_{\tau,D}(\cdot;\bar{x}) : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ defined as

$$h_{\tau,D}(z;\bar{x}) := f(\bar{x}) + \langle \nabla f(\bar{x}), z - \bar{x}\rangle + \frac{1}{2\tau}\|z - \bar{x}\|_D^2 + g(z), \quad \forall\, z \in \mathcal{H}. \tag{9}$$

Since g is strongly convex with respect to the D-norm with modulus $\mu_{g,D} \leq \mu_g/\eta_{sup}$ (see Lemma 2.1), it is easy to show that $h_{\tau,D}(\cdot;\bar{x})$ is also strongly convex with respect to the D-norm with modulus $\frac{1}{\tau}+\mu_{g,D}$. Consequently, $h_{\tau,D}(\cdot;\bar{x})$ has a unique minimizer. Then, the proximal–gradient point $\hat{x}$ with parameters τ and D is given by

$$\hat{x} := \mathrm{prox}_{\tau g}^D(\bar{x} - \tau D^{-1}\nabla f(\bar{x})) = \operatorname*{argmin}_{z\in\mathcal{H}} h_{\tau,D}(z;\bar{x}). \tag{10}$$

For our purposes, it will be convenient to consider a suitable approximation of the point $\hat{x}$ in (10) computed by means of a fixed positive tolerance parameter.

Definition 2.1 *Given $\bar{x} \in Y$, $\tau > 0$, $D \in \mathcal{D}_{\eta_{inf}}^{\eta_{sup}}$ and $\epsilon \geq 0$, we say that a point $\tilde{x} \in dom(g)$ is an $\epsilon-$approximation of the proximal–gradient point $\hat{x}$ and write $\tilde{x} \approx_\epsilon \hat{x}$ if*

$$h_{\tau,D}(\tilde{x};\bar{x}) - h_{\tau,D}(\hat{x};\bar{x}) \leq \epsilon. \tag{11}$$

We are now ready to state and prove the promised generalized descent inequality, adapted to the scaled and inexact framework introduced above.

Lemma 2.3 *Given a point $\bar{x} \in Y$, $\tau > 0$, $D \in \mathcal{D}_{\eta_{inf}}^{\eta_{sup}}$, $\epsilon \geq 0$ and a point $\tilde{x} \approx_\epsilon \hat{x}$, if $\mu_{g,D} \leq \frac{\mu_g}{\eta_{sup}}$ and $\mu_{f,D} \leq \frac{\mu_f}{\eta_{sup}}$, then the following inequality holds*

$$\begin{aligned} F(\tilde{x}) + (1+\tau\mu_{g,D})\frac{\|x-\tilde{x}\|_D^2}{2\tau} + \left(\frac{\|\tilde{x}-\bar{x}\|_D^2}{2\tau} - \mathbb{D}_f(\tilde{x},\bar{x})\right) \\ \leq F(x) + (1-\tau\mu_{f,D})\frac{\|x-\bar{x}\|_D^2}{2\tau} + \epsilon + \frac{\sqrt{2\epsilon\tau(1+\tau\mu_{g,D})}}{\tau}\|x-\tilde{x}\|_D, \quad \forall\ x \in \mathcal{H}. \end{aligned} \tag{12}$$

Proof. Let $\mu_{h,D} = \frac{1}{\tau} + \mu_{g,D}$ be the strong convexity modulus of $h_{\tau,D}(\cdot;\bar{x})$ with respect to the $D-$norm. Combining the strong convexity of $h_{\tau,D}(\cdot;\bar{x})$ with the optimality of the proximal–gradient point $\hat{x}$, we obtain that

$$h_{\tau,D}(x;\bar{x}) \geq h_{\tau,D}(\hat{x};\bar{x}) + \frac{\mu_{h,D}}{2}\|x-\hat{x}\|_D^2, \quad \forall x \in \mathcal{H}. \tag{13}$$

Choosing $x = \tilde{x}$ above and combining it with (11) leads to

$$\|\tilde{x}-\hat{x}\|_D^2 \leq \frac{2}{\mu_{h,D}}\epsilon. \tag{14}$$

Starting again from (13), we can thus write the following chain of inequalities

$$\begin{aligned} h_{\tau,D}(x;\bar{x}) &\geq h_{\tau,D}(\hat{x};\bar{x}) + \frac{\mu_{h,D}}{2}\|x-\hat{x}\|_D^2 = h_{\tau,D}(\tilde{x};\bar{x}) + h_{\tau,D}(\hat{x};\bar{x}) - h_{\tau,D}(\tilde{x};\bar{x}) + \frac{\mu_{h,D}}{2}\|x-\hat{x}\|_D^2 \\ &\geq h_{\tau,D}(\tilde{x};\bar{x}) - \epsilon + \frac{\mu_{h,D}}{2}\|x-\tilde{x}\|_D^2 + \frac{\mu_{h,D}}{2}\|\tilde{x}-\hat{x}\|_D^2 + \mu_{h,D}\langle D(x-\tilde{x}), \tilde{x}-\hat{x}\rangle \\ &\geq h_{\tau,D}(\tilde{x};\bar{x}) + \frac{\mu_{h,D}}{2}\|x-\tilde{x}\|_D^2 - \epsilon - \mu_{h,D}\sqrt{\frac{2\epsilon}{\mu_{h,D}}}\|x-\tilde{x}\|_D, \quad \forall x \in \mathcal{H} \end{aligned}$$

where the third inequality follows from the definition of $\epsilon-$approximation (11) and the three-point-equality $\|a-c\|_D^2 = \|a-b\|_D^2 + \|b-c\|_D^2 + 2\langle D(a-b), b-c\rangle$ and the fourth one from the Cauchy-Schwarz inequality, combined with (14) and the non-negativity of $\|\tilde{x}-\hat{x}\|_D^2$. We deduce

$$h_{\tau,D}(x;\bar{x}) \geq h_{\tau,D}(\tilde{x};\bar{x}) + \frac{\mu_{h,D}}{2}\|\tilde{x}-x\|^2 - \epsilon - \mu_{h,D}\sqrt{\frac{2\epsilon}{\mu_{h,D}}}\|x-\tilde{x}\|_D, \quad \forall x \in \mathcal{H}. \tag{15}$$

Then, inequality (12) follows from noticing that, for all $x \in \mathcal{H}$, we have

$$
\begin{aligned}
&F(x) + (1 - \tau\mu_{f,D})\frac{\|x - \bar{x}\|_D^2}{2\tau} \geq g(x) + f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x}\rangle + \frac{\|x - \bar{x}\|_D^2}{2\tau} = h_{\tau,D}(x;\bar{x}) \\
&\overset{(15)}{\geq} h_{\tau,D}(\tilde{x};\bar{x}) + \frac{\mu_{h,D}}{2}\|\tilde{x} - x\|^2 - \epsilon - \mu_{h,D}\sqrt{\frac{2\epsilon}{\mu_{h,D}}}\|x - \tilde{x}\|_D \\
&= g(\tilde{x}) + f(\bar{x}) + \langle \nabla f(\bar{x}), \tilde{x} - \bar{x}\rangle + \frac{\|\tilde{x} - \bar{x}\|_D^2}{2\tau} + \frac{\tau\mu_{g,D} + 1}{2\tau}\|\tilde{x} - x\|_D^2 - \epsilon - \mu_{h,D}\sqrt{\frac{2\epsilon}{\mu_{h,D}}}\|x - \tilde{x}\|_D \\
&= F(\tilde{x}) + \frac{\|\tilde{x} - \bar{x}\|_D^2}{2\tau} - \mathbb{D}_f(\tilde{x},\bar{x}) + \frac{\tau\mu_{g,D} + 1}{2\tau}\|\tilde{x} - x\|_D^2 - \epsilon - \sqrt{2\epsilon\left(\frac{1}{\tau} + \mu_{g,D}\right)}\|x - \tilde{x}\|_D,
\end{aligned}
$$

where the first inequality is due to Lemma 2.1, the first equality follows from the definition of $h_{\tau,D}$ in (9), whereas the last two equalities follow from the definitions of $h_{\tau,D}$, $\mu_{h,D}$, and F in (6). Rearranging the terms, we get the thesis. □

3 SAGE-FISTA: description and convergence analysis

In the following, we describe a generalized version of the FISTA algorithm proposed in [4], which is suited for solving the (possibly strongly) convex problem (6) by means of an appropriate scaled and inexact inertial forward-backward splitting. Our proposed algorithm, named SAGE-FISTA (Scaled Adaptive GEneralized FISTA), is obtained by incorporating in the iterative scheme (3) the following features:

- the use of a variable non-Euclidean metric in the computation of the proximal–gradient point, which is induced along the iterations by the elements of a sequence of linear, bounded and self-adjoint positive operators $\{D_k\}_{k\in\mathbb{N}}$; such operators are typically chosen in order to capture some second order information of the differentiable part f at the current iterate $x^{(k)}$ (see e.g. [9, 11, 12, 28]);
- the approximate computation of the proximal-gradient point according to the notion of ϵ -approximation given in Definition 2.1; this feature has already been introduced in other scaled FISTA-like algorithms available in the literature, see e.g. [11, 24];
- a non-monotone backtracking strategy similar to the one considered in [14], which allows for possible increasing of the step-size τ_{k+1} at each iteration; differently from monotone Armijo-type approaches, this strategy is particularly helpful when the initial τ_0 is chosen to be extremely small, which corresponds to a (pessimistic) large estimate L_0 of the Lipschitz constant value L_f;
- a novel selection rule for the inertial parameter β_{k+1}, which depends on both the strong convexity moduli μ_f, μ_g and the bounds on the spectrum of the operators D_{k+1}.

The proposed SAGE-FISTA is fully reported in Algorithm 1.

At each iteration $k \geq 0$, the preliminary step of SAGE-FISTA is the choice of both the linear, bounded and self-adjoint operator $D_{k+1} \in \mathcal{D}_{\eta_{inf}^k}^{\eta_{sup}^k}$ and the tentative step-size $\tau_{k+1}^0 = \tau_k/\delta$. If $\delta < 1$, then an adaptive increasing of τ_k is allowed (see [14, 23] for analogous strategies applied to

GFISTA and [38] for FISTA), while if $\delta = 1$, a classical Armijo-type backtracking is performed (as in [4] for FISTA). Then, an inner backtracking procedure computing the next iterate $x^{(k+1)}$ follows. At each inner backtracking iteration $i \geq 0$, the iterates τ'_{k+1}, t_{k+1} are first updated (STEPS 2–3); such quantities depend not only on the scaled quantities $\mu_{f,k+1} = \mu_f/\eta^k_{sup}$, $\mu_{g,k+1} = \mu_g/\eta^k_{sup}$ and $\mu_{k+1} = \mu_{f,k+1} + \mu_{g,k+1}$, but also on the tentative step-size τ_{k+1}. Then, the inertial parameter β_{k+1} and the projected inertial point $y^{(k+1)}$ are computed (STEPS 4–5). The projection of the inertial point $x^{(k)} + \beta_k(x^{(k)} - x^{(k-1)})$ onto the set Y is required whenever $\mathrm{dom}\,(f)$ is a proper subset of the entire space $\mathcal{H}$, in order to keep the iterate $y^{(k+1)}$ inside the domain and make the evaluation of $f(y^{(k+1)})$ and $\nabla f(y^{(k+1)})$ at STEP 6 well-defined. Although the projection onto Y can usually be solved only approximately in the general case, it becomes trivial in certain practical settings, for instance when D_k is a diagonal positive definite matrix and the domain is the non-negative orthant, i.e. $\mathrm{dom}(f) = \{x \in \mathbb{R}^n : \ x_i \geq 0, \ i = 1, \ldots, n\}$, as it is the case of the Kullback-Leibler divergence (see Section 4 and references therein). Finally, an approximated point $x^{(k+1)}$ such that $x^{(k+1)} \approx_{\epsilon_{k+1}} \mathrm{prox}^{D_{k+1}}_{\tau_{k+1}g}(y^{(k+1)} - \tau_{k+1}D^{-1}_{k+1}\nabla f(y^{(k+1)}))$ is computed. For each backtracking iteration, a check on the condition

$$\mathbb{D}_f(x^{(k+1)}, y^{(k+1)}) < \frac{\|x^{(k+1)} - y^{(k+1)}\|^2_{D_{k+1}}}{2\tau_{k+1}} \tag{16}$$

is performed (STEP 6). If such condition is not satisfied, then the step-size is reduced by a factor ρ (STEP 1) and STEPS 2–6 are performed until the condition is satisfied. Note that this procedure ends in a finite number of steps, thanks to Lemma 2.2. Furthermore, reasoning as in [33, Lemma 4], we can provide an upper bound on the total number of backtracking iterations performed by the end of the $k-$th iteration. In particular, denoting with i_j the number of backtracking iterations at step $j = 0, \ldots, k$, we have

$$\sum_{j=0}^{k} i_j \leq \left(1 + \frac{\ln(\delta)}{\ln(\rho)}\right)(k+1) + \frac{1}{\ln(\rho)}\left(\ln\left(\frac{\delta\tau_0 L_f}{\rho}\right)\right)_+ .$$

Note that such an upper bound depends solely on k, the backtracking parameters ρ, δ, the initial step-size τ_0, and the Lipschitz constant L_f.

When $\eta^k_{inf} = \eta^k_{sup} = 1$, $D_k \equiv \mathcal{I}$ and $Y = \mathcal{H}$, Algorithm 1 reduces to the Generalised FISTA (GFISTA) algorithm studied in [14]. The presence of the strong convexity moduli μ_f, μ_g in the update rule of the inertial parameter β_{k+1} allows to prove a linear convergence rate for the function values also when an adaptive backtracking strategy is adopted, see [14, Theorem 4.6].

When $D_k \neq \mathcal{I}$, Algorithm 1 can be considered as a scaled version of GFISTA where the inexact computation of the proximal-gradient point is explicitly taken into account in STEP 6. Note that the introduction of the positive operator D_k further imposes a modification of the parameter t_{k+1} in STEP 3. However, linear convergence rates still hold, as we will rigorously show in Section 3.

Finally, note that when $\mu = \mu_f + \mu_g = 0$, $\eta^k_{sup} \equiv \eta > 0$ and $\delta = 1$, STEP 3 reduces to $t_{k+1} = \frac{1+\sqrt{1+4\frac{\tau_k}{\tau_{k+1}}t_k^2}}{2}$, and Algorithm 1 becomes the inexact scaled forward-backward extrapolation method equipped with Armijo-type backtracking proposed in [11]. In this case, standard $\mathcal{O}(1/k^2)$ convergence rates can be proved, coherently with the result obtained in [11, Theorem 3.1].

Remark 3.1 *We introduce the quantity*

$$q_k := \mu_k \tau'_k = \frac{\mu_k \tau_k}{1 + \tau_k \mu_{g,k}}, \tag{17}$$

Algorithm 1 Scaled Adaptive GEneralized FISTA (SAGE-FISTA)

Parameters: $\rho \in (0,1)$, $\delta \in (0,1]$, $\{\eta_{sup}^k\}_{k\in\mathbb{N}}$ s.t. $0 < \eta_{inf} \leq \eta_{sup}^k \leq \eta_{sup}$, $\{\mu_{f,k}\}_{k\in\mathbb{N}}$, $\{\mu_{g,k}\}_{k\in\mathbb{N}}$, $\{\mu_k\}_{k\in\mathbb{N}}$ s.t. $\mu_{f,k} = \frac{\mu_f}{\eta_{sup}^k}$, $\mu_{g,k} = \frac{\mu_g}{\eta_{sup}^k}$, $\mu_k = \mu_{f,k} + \mu_{g,k}$.

Initialization: $\tau_0 > 0$ s.t. $0 \leq \tau_0\mu_{f,0} < 1$, $\tau_0' = \frac{\tau_0}{1+\tau_0\mu_{g,0}}$, $x^{(-1)} \in \mathcal{H}$, $x^{(0)} = x^{(-1)}$, and $t_0 > 0$ s.t. $1 \leq t_0 \leq \frac{1}{\sqrt{\mu_0\tau_0'}}$.

For $k = 0, 1, 2, \ldots$

Choose $D_{k+1} \in \mathcal{D}_{\eta_{inf}}^{\eta_{sup}^k}$ and set $\tau_{k+1}^0 = \frac{\tau_k}{\delta}$.

For $i = 0, 1, \ldots$ repeat

Step 1. Set $\tau_{k+1} = \rho^i \tau_{k+1}^0$.

Step 2. Set $\tau_{k+1}' = \frac{\tau_{k+1}}{1+\tau_{k+1}\mu_{g,k+1}}$.

Step 3. Set $t_{k+1} = \frac{1-\mu_k\tau_k' t_k^2 + \sqrt{(1-\mu_k\tau_k' t_k^2)^2 + 4\frac{\eta_{sup}^{k+1}\tau_k'}{\eta_{sup}^k\tau_{k+1}'}t_k^2}}{2}$.

Step 4. Set $\beta_{k+1} = \left(\frac{t_k - 1}{t_{k+1}}\right)\frac{1+\tau_{k+1}\mu_{g,k+1} - t_{k+1}\tau_{k+1}\mu_{k+1}}{1-\tau_{k+1}\mu_{f,k+1}}$.

Step 5. Set $y^{(k+1)} = P_{Y,D_{k+1}}(x^{(k)} + \beta_{k+1}(x^{(k)} - x^{(k-1)}))$.

Step 6. Set $\hat{x}^{(k+1)} = \text{prox}_{\tau_{k+1}g}^{D_{k+1}}(y^{(k+1)} - \tau_{k+1}D_{k+1}^{-1}\nabla f(y^{(k+1)}))$, choose $\epsilon_{k+1} \geq 0$ and compute $x^{(k+1)} \in \text{dom}(g)$ such that

$$x^{(k+1)} \approx_{\epsilon_{k+1}} \hat{x}^{(k+1)}.$$

until $\mathbb{D}_f(x^{(k+1)}, y^{(k+1)}) < \frac{\|x^{(k+1)} - y^{(k+1)}\|_{D_{k+1}}^2}{2\tau_{k+1}}$.

End

and observe that

$$0 \leq q_k < 1, \quad t_k \geq 1, \quad \forall\ k \geq 0.$$

Indeed q_k is nonnegative due to the nonnegativity of μ_k, μ_g, τ_k, and combining the $\mu_{f,k}-$strong convexity of f with respect to the D_k-norm with (16)*, we have*

$$\frac{\mu_{f,k}}{2}\|x^{(k)} - y^{(k)}\|_{D_k}^2 \leq \mathbb{D}_f(x^{(k)}, y^{(k)}) < \frac{\|x^{(k)} - y^{(k)}\|_{D_k}^2}{2\tau_k},$$

which implies $\mu_{f,k}\tau_k < 1$ and thus $q_k < 1$. Furthermore, in the case when $\mu = \mu_f + \mu_g = 0$, then $t_k = \frac{1+\sqrt{1+4\frac{\tau_{k-1}}{\tau_k}t_{k-1}^2}}{2} \geq 1$, otherwise we can rewrite t_k as

$$t_k = \frac{1 - q_{k-1}t_{k-1}^2 + \sqrt{\left(1 - q_{k-1}t_{k-1}^2\right)^2 + 4\frac{q_{k-1}}{q_k}t_{k-1}^2}}{2}$$

and then proceeding as in [14, Lemma 4.4] we have

$$t_k \geq \frac{1 - q_{k-1}t_{k-1}^2 + \sqrt{\left(1 - q_{k-1}t_{k-1}^2\right)^2 + 4q_{k-1}t_{k-1}^2}}{2} = \frac{1 - q_{k-1}t_{k-1}^2 + \sqrt{\left(1 + q_{k-1}t_{k-1}^2\right)^2}}{2} = 1.$$

Remark 3.2 *The update rule for the sequence* $\{t_k\}_k$ *is obtained by imposing the following condition at each iteration:*

$$\tau'_{k+1} t_{k+1}(t_{k+1}-1) = \frac{\eta_{sup}^{k+1}}{\eta_{sup}^{k}} \omega_{k+1} \tau'_k t_k^2, \quad \forall\ k \geq 0 \tag{18}$$

where we have set:

$$\omega_k := 1 - t_k q_k = \frac{1 + \tau_k \mu_{g,k} - \tau_k t_k \mu_k}{1 + \tau_k \mu_{g,k}}. \tag{19}$$

Indeed, note that t_{k+1} *is defined as the solution of the following second-degree equation*

$$t^2 - (1 - \mu_k \tau'_k t_k^2)t - \frac{\eta_{sup}^{k+1} \tau'_k}{\eta_{sup}^{k} \tau'_{k+1}} t_k^2 = 0 \quad \Leftrightarrow \quad \tau'_{k+1} t(t-1) = \frac{\eta_{sup}^{k+1}}{\eta_{sup}^{k}} (1 - \tau'_{k+1} \mu_{k+1} t) \tau'_k t_k^2,$$

where the equivalence follows by multiplying by τ'_{k+1} *and noting that* $\mu_k = \mu_{k+1}(\eta_{sup}^{k+1}/\eta_{sup}^{k})$. *Then, choosing* $t = t_{k+1}$ *and recalling the definitions of* q_k *in* (17) *and* ω_k *in* (19) *leads to* (18).

We conclude this section with the following technical result holding for the sequences $\{q_k t_k^2\}_{k\in\mathbb{N}}$ and $\{q_k t_k\}_{k\in\mathbb{N}}$, which can be proved by proceeding exactly as in [16, p. 294].

Lemma 3.1 *If* $0 \leq t_0 \leq \frac{1}{\sqrt{q_0}}$ *(as assumed in Algorithm 1), then it follows that*

$$q_k t_k^2 \leq 1 \quad \text{and} \quad q_k t_k < 1, \quad \forall\ k \geq 0. \tag{20}$$

3.1 Convergence rates

We now show that SAGE-FISTA enjoys a convergence rate in function values that is either $R-$linear or proportional to $1/k^2$, depending on whether the strong convexity modulus $\mu = \mu_f + \mu_g$ is positive or null. A couple of technical results holding for nonnegative sequences are needed in order to deal with ϵ_k-approximations and scaling operators.

Lemma 3.2 *[39, Lemma 1] Let* $\{p_k\}_{k\in\mathbb{N}}$, $\{q_k\}_{k\in\mathbb{N}}$, $\{\lambda_k\}_{k\in\mathbb{N}}$ *be sequences of real nonnegative numbers, with* $\{q_k\}_{k\in\mathbb{N}}$ *being a monotone nondecreasing sequence. Then the following result holds:*

$$p_k^2 \leq q_k + \sum_{i=1}^{k} \lambda_i p_i, \quad \forall\ k \geq 1 \quad \Rightarrow \quad p_k \leq \frac{1}{2}\sum_{i=1}^{k} \lambda_i + \left(q_k + \left(\frac{1}{2} \sum_{i=1}^{k} \lambda_i \right)^2 \right)^{\frac{1}{2}}, \quad \forall\ k \geq 1.$$

Lemma 3.3 *[36, p. 44, Lemma 2] Let* $\{p_k\}_{k\in\mathbb{N}}$, $\{\zeta_k\}_{k\in\mathbb{N}}$ *be sequences of real nonnegative numbers such that* $\sum_{k=0}^{\infty} \zeta_k < \infty$ *and the inequality* $p_{k+1} \leq (1+\zeta_k) p_k$ *holds for all* $k \geq 0$. *Then* $\{p_k\}_{k\in\mathbb{N}}$ *is a convergent sequence.*

Furthermore, we make the following assumption on the sequence of operators $\{D_k\}_{k\in\mathbb{N}}$ and the sequence of upper bounding values $\{\eta_{sup}^{k}\}_{k\in\mathbb{N}}$.

Assumption 1 *There exists a sequence of real nonnegative numbers* $\{\gamma_k\}_{k\in\mathbb{N}}$ *s.t.* $\sum_{k=0}^{\infty} \gamma_k < \infty$ *and, for all* $k \geq 0$, *the following conditions hold*

$$D_{k+1} \preceq (1+\gamma_{k+1}) D_k, \tag{21}$$

$$\frac{\eta_{sup}^{k+1}}{\eta_{sup}^{k}} \leq 1 + \gamma_{k+1}. \tag{22}$$

Remark 3.3 *Condition* (21) *has been often used to prove the convergence of variable metric FB methods in the convex setting, see for instance [8, 10, 11, 18]. It is easy to see that* (21) *implies* (22) *for the particular choice* $\eta_{sup}^k = \|D_k^{\frac{1}{2}}\|^2$, *with* $D_k^{\frac{1}{2}}$ *being the square root operator applied to* D_k; *when* $\|\cdot\| = \|\cdot\|_2$, *this amounts to choosing* $\eta_{sup}^k = \|D_k\|_2$, *i.e. the spectral radius of* D_k. *However, in general, we need to impose condition* (22) *explicitly. Let us mention two particular examples of sequences* $\{D_k\}_{k\in\mathbb{N}}$ *and* $\{\eta_{sup}^k\}_{k\in\mathbb{N}}$ *satisfying Assumption 1, which we will use in the following numerical experiments.*

- *Suppose* $D_k \equiv D$ *and* $\eta_{sup}^k \equiv \eta > 0$, *with* $D \in \mathcal{D}_{\eta_{inf}}^{\eta}$. *Then Assumption 1 holds with* $\gamma_k \equiv 0$.

- *Suppose* $D_k \in \mathcal{D}_{\eta_{inf}^k}^{\eta_{sup}^k}$ *for all* $k \geq 0$, *and the upper and lower bounds converge to the same positive value at a sufficiently fast rate, that is*

$$\eta_{inf}^k = \eta - \nu_{inf}^k, \quad \eta_{sup}^k = \eta + \nu_{sup}^k, \tag{23}$$

where $0 \leq \nu_{inf}^k < \eta$, $\nu_{sup}^k \geq 0$, *and* $\sum_{k=0}^{\infty} \nu_{inf}^k < \infty$, $\sum_{k=0}^{\infty} \nu_{sup}^k < \infty$. *In this case, recalling* (5), *we can write*

$$\|x\|_{D_{k+1}}^2 \leq \eta_{sup}^{k+1}\|x\|^2 = \frac{\eta_{sup}^{k+1}}{\eta_{inf}^k}\eta_{inf}^k\|x\|^2 \leq \frac{\eta_{sup}^{k+1}}{\eta_{inf}^k}\|x\|_{D_k}^2, \quad \forall\, x \in \mathcal{H}. \tag{24}$$

Thanks to (23), *the factor multiplying* $\|x\|_{D_k}^2$ *can be rewritten as* $\frac{\eta_{sup}^{k+1}}{\eta_{inf}^k} = 1 + \gamma_{k+1}$, *with* $\gamma_{k+1} = \frac{\nu_{sup}^{k+1}+\nu_{inf}^k}{\eta-\nu_{inf}^k}$. *Since* $\gamma_{k+1} \leq (\eta - \max_k \nu_{inf}^k)^{-1}(\nu_{sup}^{k+1}+\nu_{inf}^k)$ *and the sequences* $\{\nu_{inf}^k\}_{k\in\mathbb{N}}$, $\{\nu_{sup}^k\}_{k\in\mathbb{N}}$ *are summable, we can conclude that condition* (21) *is satisfied. Furthermore, we also have*

$$\frac{\eta_{sup}^{k+1}}{\eta_{sup}^k} = \frac{\eta_{sup}^{k+1}}{\eta_{inf}^k} \cdot \frac{\eta_{inf}^k}{\eta_{sup}^k} \leq \frac{\eta_{sup}^{k+1}}{\eta_{inf}^k} = 1 + \gamma_{k+1},$$

that is condition (22) *also holds.*
Note that, if $\mathcal{H} = \mathbb{R}^n$, $\|\cdot\| = \|\cdot\|_2$, *and* $\{D_k\}_{k\in\mathbb{N}}$ *are diagonal matrices, then we can always impose condition* (23) *by constraining the diagonal elements of* D_k *in the interval* $[\eta - \nu_{inf}^k, \eta + \nu_{sup}^k]$ *(see Section 4 for a practical implementation). By doing so, we progressively "squeeze" the scaling matrices to a multiple of the identity matrix as the iterations increase. This allows us to employ very well-known techniques for computing (diagonal) scaling matrices, such as Majorization–Minimization or Split–Gradient strategies [8, 9, 28], provided that their elements are artificially thresholded from below and above.*

On the contrary, conditions (21)-(22) *are not suited for Quasi-Newton variable metrics, which aim at estimating a certain Hessian matrix through low-rank updates (see e.g. [6]), rather than converging to a multiple of the identity matrix (as* (21)-(22) *require). We remark that FB methods exploiting the combination of Quasi-Newton metrics, inertial steps and inexact proximal evaluations have already been devised in the literature, for instance in [24, 26]. In [26], the typical* $\mathcal{O}(1/k^2)$ *convergence rate is achieved by requiring the monotonicity condition* $0 \preceq D_k \preceq D_{k-1}$ *on the scaling matrices; in [24], the authors prove the same convergence rate for a similar Accelerated Proximal quasi-Newton Algorithm (APQNA) by relaxing the monotonicity assumption on the metric. However, as highlighted in [24, p. 612], it is not clear whether these theoretical assumptions are compatible with Quasi-Newton approaches such as L-BFGS, and the numerical experiments in [24] suggest that different techniques other than standard Nesterov-type inertia*

might be required to obtain satisfactory acceleration for proximal Quasi-Newton methods. Other works [5, 6] have proved convergence results for Quasi-Newton (non-inertial) forward–backward methods in the strongly convex setting, by developing a suitable proximable calculus that allows to efficiently evaluate (or exactly compute, in certain cases) the proximal operator with respect to Quasi–Newton variable metrics. By contrast, in this work we pursue a complementary approach, which is tailored for specific diagonal metrics (e.g. the ones obtained with Majorization–Minimization or Split–Gradient strategies).

We now provide a key descent inequality that holds for the iterates of SAGE-FISTA Algorithm 1, in the same spirit of the one employed in [14, p. 8]. The result follows by combining Lemma 2.3 with the backtracking condition of Algorithm 1 and Assumption 1.

Lemma 3.4 *Suppose that $F = f + g$ is $\mu-$strongly convex with $\mu \geq 0$ and Assumption 1 holds. Then, for all $k \geq 0$, we have*

$$\begin{aligned}
&\tau'_{k+1}t_{k+1}^2(F(x^{(k+1)}) - F(x^*)) + \frac{1}{2}\|x^* - x^{(k+1)} - (t_{k+1}-1)(x^{(k+1)} - x^{(k)})\|^2_{D_{k+1}} \\
&\leq \omega_{k+1}(1+\gamma_{k+1})\left(\tau'_k t_k^2(F(x^{(k)}) - F(x^*)) + \frac{1}{2}\|x^* - x^{(k)} - (t_k-1)(x^{(k)} - x^{(k-1)})\|^2_{D_k}\right) \\
&+ \tau'_{k+1}t_{k+1}^2\epsilon_{k+1} + t_{k+1}\sqrt{2\epsilon_{k+1}\tau'_{k+1}}\|x^* - x^{(k+1)} - (t_{k+1}-1)(x^{(k+1)} - x^{(k)})\|_{D_{k+1}}.
\end{aligned}$$

Proof. We apply Lemma 2.3 with $\tau = \tau_{k+1}$, $\bar{x} = y^{(k+1)}$, $\tilde{x} = x^{(k+1)}$, $\epsilon = \epsilon_{k+1}$, $D = D_{k+1}$, $\mu_{f,D} = \mu_{f,k+1}$, $\mu_{g,D} = \mu_{g,k+1}$ and combine it with the backtracking condition (16) so as to obtain:

$$\begin{aligned}
F(x^{(k+1)}) + (1+\tau_{k+1}\mu_{g,k+1})\frac{\|x - x^{(k+1)}\|^2_{D_{k+1}}}{2\tau_{k+1}} &\leq F(x) + (1-\tau_{k+1}\mu_{f,k+1})\frac{\|x - y^{(k+1)}\|^2_{D_{k+1}}}{2\tau_{k+1}} \\
&+ \epsilon_{k+1} + \frac{\sqrt{2\epsilon_{k+1}\tau_{k+1}(1+\tau_{k+1}\mu_{g,k+1})}}{\tau_{k+1}}\|x - x^{(k+1)}\|_{D_{k+1}}.
\end{aligned}$$

Let us now define $x = \frac{(t_{k+1}-1)x^{(k)}+x^*}{t_{k+1}}$ with $x^{(k)}$ being the $k-$th iterate of the algorithm and $x^* \in \mathcal{H}$ a solution of (6), and set $\bar{y}^{(k+1)} = x^{(k)} + \beta_{k+1}(x^{(k)} - x^{(k-1)})$. Since the operator $P_{Y,D_{k+1}}$ is firmly non-expansive [11, Remark 2.2] and $x \in Y$, we have $\|x - y^{(k+1)}\|^2_{D_{k+1}} \leq \|x - \bar{y}^{(k+1)}\|^2_{D_{k+1}}$. Then, the previous inequality becomes

$$\begin{aligned}
F(x^{(k+1)}) + (1+\tau_{k+1}\mu_{g,k+1})\frac{\|(t_{k+1}-1)x^{(k)} + x^* - t_{k+1}x^{(k+1)}\|^2_{D_{k+1}}}{2\tau_{k+1}t_{k+1}^2} &\leq F\left(\frac{(t_{k+1}-1)x^{(k)} + x^*}{t_{k+1}}\right) \\
&+ (1-\tau_{k+1}\mu_{f,k+1})\frac{\|(t_{k+1}-1)x^{(k)} + x^* - t_{k+1}\bar{y}^{(k+1)}\|^2_{D_{k+1}}}{2\tau_{k+1}t_{k+1}^2} \\
&+ \epsilon_{k+1} + \frac{\sqrt{2\epsilon_{k+1}\tau_{k+1}(1+\tau_{k+1}\mu_{g,k+1})}}{\tau_{k+1}t_{k+1}}\|(t_{k+1}-1)x^{(k)} + x^* - t_{k+1}x^{(k+1)}\|_{D_{k+1}}. \quad (25)
\end{aligned}$$

We now multiply both sides of (25) by t_{k+1}^2, apply the $\mu_{k+1}-$strong convexity of F with respect to the $D_{k+1}-$norm on the right-hand side (the point $\frac{(t_{k+1}-1)x^{(k)}+x^*}{t_{k+1}}$ is a convex combination of

$x^{(k)}$ and x^*) and subtract the term $t_{k+1}^2 F(x^*)$ to both sides, thus obtaining

$$\begin{aligned}
&t_{k+1}^2(F(x^{(k+1)}) - F(x^*)) + \frac{1+\tau_{k+1}\mu_{g,k+1}}{2\tau_{k+1}}\|(t_{k+1}-1)x^{(k)} + x^* - t_{k+1}x^{(k+1)}\|_{D_{k+1}}^2 \\
&\quad \leq t_{k+1}(t_{k+1}-1)(F(x^{(k)}) - F(x^*)) - \frac{\mu_{k+1}(t_{k+1}-1)}{2}\|x^{(k)} - x^*\|_{D_{k+1}}^2 \\
&\quad + (1-\tau_{k+1}\mu_{f,k+1})\frac{\|(t_{k+1}-1)x^{(k)} + x^* - t_{k+1}\bar{y}^{(k+1)}\|_{D_{k+1}}^2}{2\tau_{k+1}} \\
&\quad + t_{k+1}^2\epsilon_{k+1} + t_{k+1}\frac{\sqrt{2\epsilon_{k+1}\tau_{k+1}(1+\tau_{k+1}\mu_{g,k+1})}}{\tau_{k+1}}\|(t_{k+1}-1)x^{(k)} + x^* - t_{k+1}x^{(k+1)}\|_{D_{k+1}}.
\end{aligned} \tag{26}$$

By proceeding similarly as in [16, p. 292], we rewrite part of the right-hand side of inequality (26) as follows:

$$\begin{aligned}
&-\frac{\mu_{k+1}(t_{k+1}-1)}{2}\|x^{(k)} - x^*\|_{D_{k+1}}^2 + (1-\tau_{k+1}\mu_{f,k+1})\frac{\|(t_{k+1}-1)x^{(k)} + x^* - t_{k+1}\bar{y}^{(k+1)}\|_{D_{k+1}}^2}{2\tau_{k+1}} \\
&= \frac{1-\tau_{k+1}\mu_{f,k+1} - \mu_{k+1}\tau_{k+1}(t_{k+1}-1)}{2\tau_{k+1}}\|x^{(k)} - x^*\|_{D_{k+1}}^2 \\
&+ \frac{(1-\tau_{k+1}\mu_{f,k+1})t_{k+1}}{\tau_{k+1}}\langle x^* - x^{(k)}, D_{k+1}(x^{(k)} - \bar{y}^{(k+1)})\rangle + \frac{t_{k+1}^2(1-\tau_{k+1}\mu_{f,k+1})}{2\tau_{k+1}}\|x^{(k)} - \bar{y}^{(k+1)}\|_{D_{k+1}}^2 \\
&= \frac{1+\tau_{k+1}\mu_{g,k+1} - t_{k+1}\tau_{k+1}\mu_{k+1}}{2\tau_{k+1}}\|x^{(k)} - x^*\|_{D_{k+1}}^2 \\
&+ \frac{t_{k+1}(1-\tau_{k+1}\mu_{f,k+1})(1+\tau_{k+1}\mu_{g,k+1} - t_{k+1}\tau_{k+1}\mu_{k+1})}{\tau_{k+1}}\left\langle x^* - x^{(k)}, D_{k+1}\left(\frac{x^{(k)} - \bar{y}^{(k+1)}}{1+\tau_{k+1}\mu_{g,k+1} - t_{k+1}\tau_{k+1}\mu_{k+1}}\right)\right\rangle \\
&+ \frac{1+\tau_{k+1}\mu_{g,k+1} - t_{k+1}\tau_{k+1}\mu_{k+1}}{2\tau_{k+1}}\left\|\frac{t_{k+1}(1-\tau_{k+1}\mu_{f,k+1})(x^{(k)} - \bar{y}^{(k+1)})}{1+\tau_{k+1}\mu_{g,k+1} - t_{k+1}\tau_{k+1}\mu_{k+1}}\right\|_{D_{k+1}}^2 \\
&+ \frac{t_{k+1}^2(1-\tau_{k+1}\mu_{f,k+1})}{2\tau_{k+1}}\left(1 - \frac{1-\tau_{k+1}\mu_{f,k+1}}{1+\tau_{k+1}\mu_{g,k+1} - t_{k+1}\tau_{k+1}\mu_{k+1}}\right)\|x^{(k)} - \bar{y}^{(k+1)}\|_{D_{k+1}}^2 \\
&= \frac{1+\tau_{k+1}\mu_{g,k+1} - t_{k+1}\tau_{k+1}\mu_{k+1}}{2\tau_{k+1}}\left\|x^* - x^{(k)} + \frac{t_{k+1}(1-\tau_{k+1}\mu_{f,k+1})}{1+\tau_{k+1}\mu_{g,k+1} - t_{k+1}\tau_{k+1}\mu_{k+1}}(x^{(k)} - \bar{y}^{(k+1)})\right\|_{D_{k+1}}^2 \\
&- \frac{t_{k+1}^2(t_{k+1}-1)\mu_{k+1}(1-\tau_{k+1}\mu_{f,k+1})}{2(1+\tau_{k+1}\mu_{g,k+1} - t_{k+1}\tau_{k+1}\mu_{k+1})}\|x^{(k)} - \bar{y}^{(k+1)}\|_{D_{k+1}}^2,
\end{aligned}$$

where the first equality is obtained by applying the basic relation $\|a+b\|_D^2 = \|a\|_D^2 + \|b\|_D^2 + 2\langle a, Db\rangle$ with $a = x^* - x^{(k)}$, $b = t_{k+1}(x^{(k)} - \bar{y}^{(k+1)})$, and $D = D_{k+1}$, the second one follows by recalling that $\mu_{k+1} = \mu_{f,k+1} + \mu_{g,k+1}$, and by adding and subtracting to the right-hand side of the equality the term $t_{k+1}^2(1-\tau_{k+1}\mu_{f,k+1})^2\|x^{(k)} - \bar{y}^{(k+1)}\|_{D_{k+1}}^2/(2\tau_{k+1}(1+\tau_{k+1}\mu_{g,k+1} - t_{k+1}\tau_{k+1}\mu_{k+1}))$, and the last equality is due again to the basic relation with $a = x^* - x^{(k)}$, $b = t_{k+1}(1-\tau_{k+1}\mu_{f,k+1})(x^{(k)} - \bar{y}^{(k+1)})/(1+\tau_{k+1}\mu_{g,k+1} - t_{k+1}\tau_{k+1}\mu_{k+1})$, $D = D_{k+1}$. We can

now plug this into (26), thus getting:

$$t_{k+1}^2(F(x^{(k+1)})-F(x^*))+\frac{1+\tau_{k+1}\mu_{g,k+1}}{2\tau_{k+1}}\|x^*-x^{(k+1)}-(t_{k+1}-1)(x^{(k+1)}-x^{(k)})\|_{D_{k+1}}^2$$
$$+\frac{t_{k+1}^2(t_{k+1}-1)\mu_{k+1}(1-\tau_{k+1}\mu_{f,k+1})}{2(1+\tau_{k+1}\mu_{g,k+1}-t_{k+1}\tau_{k+1}\mu_{k+1})}\|x^{(k)}-\bar{y}^{(k+1)}\|_{D_{k+1}}^2\leq t_{k+1}(t_{k+1}-1)(F(x^{(k)})-F(x^*))$$
$$+\frac{1+\tau_{k+1}\mu_{g,k+1}-t_{k+1}\tau_{k+1}\mu_{k+1}}{2\tau_{k+1}}\left\|x^*-x^{(k)}-\frac{t_{k+1}(1-\tau_{k+1}\mu_{f,k+1})}{1+\tau_{k+1}\mu_{g,k+1}-t_{k+1}\tau_{k+1}\mu_{k+1}}(\bar{y}^{(k+1)}-x^{(k)})\right\|_{D_{k+1}}^2$$
$$+t_{k+1}^2\epsilon_{k+1}+t_{k+1}\frac{\sqrt{2\epsilon_{k+1}\tau_{k+1}(1+\tau_{k+1}\mu_{g,k+1})}}{\tau_{k+1}}\|x^*-x^{(k+1)}-(t_{k+1}-1)(x^{(k+1)}-x^{(k)})\|_{D_{k+1}}.$$

Recalling the definition of τ'_{k+1} at STEP 2 of Algorithm 1 and ω_{k+1} in (19), we observe that $0<\omega_{k+1}\leq 1$ since $0\leq q_{k+1}t_{k+1}<1$ (see Lemma 3.1). Furthermore, the parameters β_{k+1} appearing in Algorithm 1 can be rewritten in terms of ω_{k+1} as

$$\beta_{k+1}=\omega_{k+1}\frac{t_k-1}{t_{k+1}}\frac{1+\tau_{k+1}\mu_{g,k+1}}{1-\tau_{k+1}\mu_{f,k+1}}. \tag{27}$$

Recalling also the definition of $\bar{y}^{(k+1)}$, plugging (19) and (27) inside the previous inequality and multiplying it by τ'_{k+1}, we deduce:

$$\begin{aligned}&\tau'_{k+1}t_{k+1}^2(F(x^{(k+1)})-F(x^*))+\frac{1}{2}\|x^*-x^{(k+1)}-(t_{k+1}-1)(x^{(k+1)}-x^{(k)})\|_{D_{k+1}}^2\\&\leq\tau'_{k+1}t_{k+1}(t_{k+1}-1)(F(x^{(k)})-F(x^*))+\frac{\omega_{k+1}}{2}\|x^*-x^{(k)}-(t_k-1)(x^{(k)}-x^{(k-1)})\|_{D_{k+1}}^2\\&+\tau'_{k+1}t_{k+1}^2\epsilon_{k+1}+t_{k+1}\sqrt{2\epsilon_{k+1}\tau'_{k+1}}\|x^*-x^{(k+1)}-(t_{k+1}-1)(x^{(k+1)}-x^{(k)})\|_{D_{k+1}}.\end{aligned} \tag{28}$$

Combining equation (18) with (28) yields

$$\begin{aligned}&\tau'_{k+1}t_{k+1}^2(F(x^{(k+1)})-F(x^*))+\frac{1}{2}\|x^*-x^{(k+1)}-(t_{k+1}-1)(x^{(k+1)}-x^{(k)})\|_{D_{k+1}}^2\\&\leq\omega_{k+1}\left(\frac{\eta_{sup}^{k+1}}{\eta_{sup}^{k}}\tau'_kt_k^2(F(x^{(k)})-F(x^*))+\frac{1}{2}\|x^*-x^{(k)}-(t_k-1)(x^{(k)}-x^{(k-1)})\|_{D_{k+1}}^2\right)\\&+\tau'_{k+1}t_{k+1}^2\epsilon_{k+1}+t_{k+1}\sqrt{2\epsilon_{k+1}\tau'_{k+1}}\|x^*-x^{(k+1)}-(t_{k+1}-1)(x^{(k+1)}-x^{(k)})\|_{D_{k+1}}\end{aligned}$$

whence the thesis follows by applying Assumption 1. □

Let us now define the sequence $\{\theta_k\}_{k\in\mathbb{N}}$ by

$$\theta_k:=\frac{\prod_{i=0}^{k}\omega_i}{\tau'_kt_k^2},\quad\forall\ k\geq 0. \tag{29}$$

The convergence rate for the function values of the iterates of Algorithm 1 can be computed by a careful study of the factor θ_k, which decays linearly if $\mu>0$ and quadratically if $\mu=0$.

Theorem 3.1 *Suppose $F=f+g$ is $\mu-$strongly convex with $\mu=\mu_f+\mu_g\geq 0$ and Assumption 1 holds. Let x^* be a solution of* (6) *and recall definitions* (19) *and* (29). *Set $u^{(0)}=x^*-x^{(0)}$, $v_0=$*

$F(x^{(0)})-F(x^*)$, $\gamma := \lim_{k\to\infty}\prod_{i=1}^{k}(1+\gamma_i) < \infty$ *and* $E_1^k := \sum_{i=0}^{k}\sqrt{\theta_{i+1}^{-1}\epsilon_{i+1}}$, $E_2^k := \sum_{i=0}^{k}\theta_{i+1}^{-1}\epsilon_{i+1}$. *Then, for all* $k \geq 0$, *the following upper bound for the function values holds:*

$$F(x^{(k+1)}) - F(x^*) \leq \gamma\theta_{k+1}\left(\frac{\|u^{(0)}\|_{D_0}}{\sqrt{2\omega_0}} + \sqrt{\frac{\tau_0' t_0^2 v_0}{\omega_0}} + 2\sqrt{\gamma}E_1^k + \sqrt{E_2^k}\right)^2. \tag{30}$$

Furthermore, the factor θ_{k+1} *can be bounded as follows:*

$$\theta_{k+1} \leq \frac{\eta_{sup}}{\eta_{inf}} \cdot \min\left\{\left(1 - \sqrt{\frac{\mu\rho}{L_f\eta_{sup} + \mu_g\rho}}\right)^{k+1}\left(\frac{1}{\tau_0} - \mu_{f,0}\right), \frac{4}{(k+2)^2}\sqrt{\frac{L_f\eta_{sup} - \rho\mu_f}{\rho\eta_{sup}}}\right\}. \tag{31}$$

Proof. The result is obtained by combining the proof in [14, Theorem 4.6] with some arguments employed in [39, Propositions 2-4]. Using the short-hand notations

$$u^{(k)} := x^* - x^{(k)} - (t_k - 1)(x^{(k)} - x^{(k-1)}), \quad v_k := F(x^{(k)}) - F(x^*),$$

we reformulate the inequality in Lemma 3.4 in the following compact form:

$$\begin{aligned}\tau_{k+1}' t_{k+1}^2 v_{k+1} + \frac{1}{2}\|u^{(k+1)}\|_{D_{k+1}}^2 &\leq \omega_{k+1}(1+\gamma_{k+1})\left(\tau_k' t_k^2 v_k + \frac{1}{2}\|u^{(k)}\|_{D_k}^2\right)\\ &\quad + \tau_{k+1}' t_{k+1}^2\epsilon_{k+1} + t_{k+1}\sqrt{2\epsilon_{k+1}\tau_{k+1}'}\|u^{(k+1)}\|_{D_{k+1}}.\end{aligned} \tag{32}$$

By recursively applying (32), we obtain

$$\begin{aligned}\tau_{k+1}' t_{k+1}^2 v_{k+1} + \frac{1}{2}\|u^{(k+1)}\|_{D_{k+1}}^2 &\leq \left(\prod_{i=1}^{k+1}\omega_i(1+\gamma_i)\right)\left(\tau_0' t_0^2 v_0 + \frac{1}{2}\|u^{(0)}\|_{D_0}^2\right)\\ &\quad + \sum_{i=0}^{k}\left(\prod_{j=i+2}^{k+1}\omega_j(1+\gamma_j)\right)\left(\tau_{i+1}' t_{i+1}^2\epsilon_{i+1} + t_{i+1}\sqrt{2\epsilon_{i+1}\tau_{i+1}'}\|u^{(i+1)}\|_{D_{i+1}}\right),\end{aligned} \tag{33}$$

where we used the notation $\prod_{j=i+2}^{k+1}\omega_j(1+\gamma_j) = 1$ whenever $i > k-1$.

Let us define the sequence $\{p_k\}_{k\in\mathbb{N}}$ with elements $p_k = \prod_{i=1}^{k}(1+\gamma_i)$ for all $k \geq 1$. We notice that $p_{k+1} = (1+\gamma_{k+1})p_k$. Recalling the summability of the sequence $\{\gamma_k\}_{k\in\mathbb{N}}$ and applying Lemma 3.3, it follows that $\{p_k\}_{k\in\mathbb{N}}$ converges. Setting $\gamma = \lim_{k\to\infty}\prod_{i=1}^{k}(1+\gamma_i) < \infty$, we deduce from (33)

$$\begin{aligned}\tau_{k+1}' t_{k+1}^2 v_{k+1} + \frac{1}{2}\|u^{(k+1)}\|_{D_{k+1}}^2 &\leq \left(\prod_{i=1}^{k+1}\omega_i\right)\left(\gamma\tau_0' t_0^2 v_0 + \frac{\gamma}{2}\|u^{(0)}\|_{D_0}^2\right)\\ &\quad + \sum_{i=0}^{k}\left(\prod_{j=i+2}^{k+1}\omega_j\right)\left(\gamma\tau_{i+1}' t_{i+1}^2\epsilon_{i+1} + \gamma t_{i+1}\sqrt{2\epsilon_{i+1}\tau_{i+1}'}\|u^{(i+1)}\|_{D_{i+1}}\right).\end{aligned}$$

Let us now define $\Omega_{k+1} := \prod_{i=0}^{k+1}\omega_i \in (0,1]$. Recalling the definition of θ_k in (29), we can rewrite the previous inequality as:

$$\begin{aligned}&\tau_{k+1}' t_{k+1}^2 v_{k+1} + \frac{1}{2}\|u^{(k+1)}\|_{D_{k+1}}^2\\ &\leq \gamma\Omega_{k+1}\left(\frac{\tau_0' t_0^2 v_0}{\omega_0} + \frac{1}{2\omega_0}\|u^{(0)}\|_{D_0}^2 + \sum_{i=0}^{k}\theta_{i+1}^{-1}\epsilon_{i+1} + \sqrt{2\theta_{i+1}^{-1}\epsilon_{i+1}}\Omega_{i+1}^{-\frac{1}{2}}\|u^{(i+1)}\|_{D_{i+1}}\right).\end{aligned} \tag{34}$$

We can now proceed as in [39, Proposition 2]. First, we discard the term $\tau'_{k+1}t_{k+1}^2 v_{k+1}$ on the left-hand side, so as to deduce the following inequality

$$\Omega_{k+1}^{-1}\|u^{(k+1)}\|_{D_{k+1}}^2 \leq 2\gamma\left(\frac{\|u^{(0)}\|_{D_0}^2 + 2\tau'_0 t_0^2 v_0}{2\omega_0} + \sum_{i=0}^{k}\theta_{i+1}^{-1}\epsilon_{i+1}\right) + \sum_{i=0}^{k} 2\gamma\sqrt{2\theta_{i+1}^{-1}\epsilon_{i+1}}\,\Omega_{i+1}^{-\frac{1}{2}}\|u^{(i+1)}\|_{D_{i+1}}.$$

An application of Lemma 3.2 to this relation leads to

$$\begin{aligned}\Omega_{k+1}^{-\frac{1}{2}}\|u^{(k+1)}\|_{D_{k+1}} &\leq \sum_{i=0}^{k}\gamma\sqrt{2\theta_{i+1}^{-1}\epsilon_{i+1}} \\ &\quad + \left(2\gamma\left(\frac{\|u^{(0)}\|_{D_0}^2 + 2\tau'_0 t_0^2 v_0}{2\omega_0} + \sum_{i=0}^{k}\theta_{i+1}^{-1}\epsilon_{i+1}\right) + \left(\sum_{i=0}^{k}\gamma\sqrt{2\theta_{i+1}^{-1}\epsilon_{i+1}}\right)^2\right)^{\frac{1}{2}}.\end{aligned}$$

Therefore, by standard inequalities on the square root of the sum of positive terms, we can deduce that for all $j = 0, \ldots, k$

$$\Omega_{j+1}^{-\frac{1}{2}}\|u^{(j+1)}\|_{D_{j+1}} \leq \sqrt{\frac{\gamma}{\omega_0}}\|u^{(0)}\|_{D_0} + \sqrt{\frac{2\gamma\tau'_0 t_0^2 v_0}{\omega_0}} + \sqrt{2\sum_{i=0}^{k}\gamma\theta_{i+1}^{-1}\epsilon_{i+1}} + 2\sum_{i=0}^{k}\gamma\sqrt{2\theta_{i+1}^{-1}\epsilon_{i+1}}. \quad (35)$$

We now go back to (34), discard the term $\frac{1}{2}\|u^{(k+1)}\|_{D_{k+1}}^2$ on the left-hand side, plug the relation (35) and write the following chain of inequalities

$$\begin{aligned}\theta_{k+1}^{-1}v_{k+1} &\leq \frac{\gamma\tau'_0 t_0^2 v_0}{\omega_0} + \frac{\gamma}{2\omega_0}\|u^{(0)}\|_{D_0}^2 + \sum_{i=0}^{k}\gamma\theta_{i+1}^{-1}\epsilon_{i+1} \\ &+ \left(2\sum_{i=0}^{k}\gamma\sqrt{\theta_{i+1}^{-1}\epsilon_{i+1}}\right)\left(\sqrt{\frac{\gamma}{2\omega_0}}\|u^{(0)}\|_{D_0} + \sqrt{\frac{\gamma\tau'_0 t_0^2 v_0}{\omega_0}} + \sqrt{\sum_{i=0}^{k}\gamma\theta_{i+1}^{-1}\epsilon_{i+1}} + 2\sum_{i=0}^{k}\gamma\sqrt{\theta_{i+1}^{-1}\epsilon_{i+1}}\right) \\ &\leq \left(\sqrt{\frac{\gamma}{2\omega_0}}\|u^{(0)}\|_{D_0} + \sqrt{\frac{\gamma\tau'_0 t_0^2 v_0}{\omega_0}} + 2\sum_{i=0}^{k}\gamma\sqrt{\theta_{i+1}^{-1}\epsilon_{i+1}} + \sqrt{\sum_{i=0}^{k}\gamma\theta_{i+1}^{-1}\epsilon_{i+1}}\right)^2.\end{aligned}$$

We thus deduce

$$v_{k+1} \leq \gamma\theta_{k+1}\left(\frac{\|u^{(0)}\|_{D_0}}{\sqrt{2\omega_0}} + \sqrt{\frac{\tau'_0 t_0^2 v_0}{\omega_0}} + 2\sum_{i=0}^{k}\sqrt{\gamma\theta_{i+1}^{-1}\epsilon_{i+1}} + \sqrt{\sum_{i=0}^{k}\theta_{i+1}^{-1}\epsilon_{i+1}}\right)^2. \quad (36)$$

At this point, proceeding similarly to [14, Theorem 4.6], we need to provide an upper bound for the factor θ_k. To this aim, we introduce the following average quantities:

$$\sqrt{\bar{L}_{k+1}} := \frac{1}{\frac{1}{k+2}\sum_{i=0}^{k+1}\frac{1}{\sqrt{L_i - \mu_{f,i}}}}, \qquad \sqrt{\bar{q}_{k+1}} := \frac{1}{k+1}\sum_{i=1}^{k+1}\sqrt{\frac{\mu_i}{L_i + \mu_{g,i}}}, \quad (37)$$

where $L_i := 1/\tau_i$. In order to show the quadratic decay of θ_k, we first prove the inequality:

$$\frac{1}{\sqrt{\eta_{sup}^k\theta_k}} \geq \frac{1}{2}\sum_{i=0}^{k}\frac{1}{\sqrt{\eta_{sup}^i\left(\frac{1}{\tau_i} - \mu_{f,i}\right)}}, \quad \forall\ k \geq 0. \quad (38)$$

The proof follows by induction: for $k = 0$ we have

$$\theta_0 = \frac{1 - \mu_0 \tau_0' t_0}{\tau_0' t_0^2} = \frac{(1 - \mu_0 \tau_0' t_0)(1 + \tau_0 \mu_{g,0})}{\tau_0 t_0^2} = \frac{1 - (t_0 - 1)\tau_0 \mu_{g,0} - \mu_{f,0} \tau_0 t_0}{\tau_0 t_0^2} \leq \frac{1}{\tau_0} - \mu_{f,0}, \quad (39)$$

which holds since, by assumption, $t_0 \geq 1$. We observe that this is nothing but (38) for $k = 0$. By induction, suppose now that (38) holds for some $k \geq 0$. Then, from (18), we deduce:

$$1 - \frac{1}{t_{k+1}} = \frac{\eta_{sup}^{k+1}}{\eta_{sup}^k} \omega_{k+1} \frac{\tau_k' t_k^2}{\tau_{k+1}' t_{k+1}^2} = \frac{\eta_{sup}^{k+1}}{\eta_{sup}^k} \frac{\theta_{k+1}}{\theta_k} \leq 1, \quad (40)$$

which implies that the sequence $\{\eta_{sup}^k \theta_k\}_{k \in \mathbb{N}}$ is non-increasing. Therefore, we have

$$\frac{1}{\sqrt{\eta_{sup}^{k+1} \theta_{k+1}}} - \frac{1}{\sqrt{\eta_{sup}^k \theta_k}} = \frac{\eta_{sup}^k \theta_k - \eta_{sup}^{k+1} \theta_{k+1}}{\sqrt{\eta_{sup}^k \eta_{sup}^{k+1} \theta_k \theta_{k+1}} \left(\sqrt{\eta_{sup}^k \theta_k} + \sqrt{\eta_{sup}^{k+1} \theta_{k+1}} \right)} \geq \frac{\eta_{sup}^k \theta_k - \eta_{sup}^{k+1} \theta_{k+1}}{2 \eta_{sup}^k \theta_k \sqrt{\eta_{sup}^{k+1} \theta_{k+1}}}.$$

Now, recalling the definitions of the parameters θ_{k+1} and ω_{k+1}, we first observe that:

$$\begin{aligned} t_{k+1} \sqrt{\eta_{sup}^{k+1} \theta_{k+1}} &= \sqrt{\frac{\eta_{sup}^{k+1}}{\tau_{k+1}'}} \prod_{i=0}^{k+1} \sqrt{\omega_i} \leq \sqrt{\frac{\eta_{sup}^{k+1} \omega_{k+1}}{\tau_{k+1}'}} = \sqrt{\eta_{sup}^{k+1} \left(\frac{1}{\tau_{k+1}'} - \mu_{k+1} t_{k+1} \right)} \\ &\leq \sqrt{\eta_{sup}^{k+1} \left(\frac{1}{\tau_{k+1}'} - \mu_{k+1} \right)} = \sqrt{\eta_{sup}^{k+1} \left(\frac{1}{\tau_{k+1}} - \mu_{f,k+1} \right)}. \end{aligned}$$

Recalling (40), this entails:

$$\frac{1}{\sqrt{\eta_{sup}^{k+1} \theta_{k+1}}} - \frac{1}{\sqrt{\eta_{sup}^k \theta_k}} \geq \frac{1}{2 t_{k+1} \sqrt{\eta_{sup}^{k+1} \theta_{k+1}}} \geq \frac{1}{2 \sqrt{\eta_{sup}^{k+1} \left(\frac{1}{\tau_{k+1}} - \mu_{f,k+1} \right)}}.$$

By induction, the previous inequality yields (38) with $k = k+1$. Recalling the definition of $\bar{L}_{k+1}$ and observing that $\eta_{inf} \leq \eta_{sup}^i \leq \eta_{sup}$ for $i = 0, \ldots, k$, we get

$$\sqrt{\theta_{k+1}} \leq \frac{2}{k+2} \sqrt{\frac{\eta_{sup}}{\eta_{inf}} \bar{L}_{k+1}}, \quad \forall\ k \geq 0. \quad (41)$$

To get the linear rate, we use (40) and (20) and deduce the following relations:

$$\theta_{k+1} = \theta_0 \prod_{i=1}^{k+1} \frac{\eta_{sup}^{i-1}}{\eta_{sup}^i} \left(1 - \frac{1}{t_i} \right) \leq \frac{\eta_{sup}^0 \theta_0}{\eta_{sup}^{k+1}} \prod_{i=1}^{k+1} (1 - \sqrt{q_i}) \leq \frac{\eta_{sup}}{\eta_{inf}} (L_0 - \mu_{f,0}) \left(1 - \sqrt{\bar{q}_{k+1}} \right)^{k+1}, \quad (42)$$

where the last inequality follows from the concavity of the logarithmic function, the definition of $\bar{q}_{k+1}$, (39) and the fact that $\eta_{inf} \leq \eta_{sup}^i \leq \eta_{sup}$ for $i = 0, \ldots, k+1$.
Finally, we note that the averaging term $\bar{L}_{k+1}$ appearing in (41) is always smaller than the actual average of the terms $(L_i - \mu_{f,i})$, since:

$$\sqrt{\bar{L}_k} \leq \frac{1}{k+1} \sum_{i=0}^{k} \sqrt{L_i - \mu_{f,i}} \leq \sqrt{\frac{1}{k+1} \sum_{i=0}^{k} (L_i - \mu_{f,i})}. \quad (43)$$

Then, since the use of the backtracking strategy at any iteration implies that $\tau_k \geq \rho/L_f$, we can deduce the following uniform bounds for the averaging terms:

$$\sqrt{\bar{L}_{k+1}} \leq \sqrt{\frac{L_f}{\rho} - \frac{\mu_f}{\eta_{sup}}}, \qquad \sqrt{\bar{q}_{k+1}} \geq \sqrt{\frac{\mu\rho}{L_f\eta_{sup} + \mu_g\rho}},$$

by which (31) follows. □

Remark 3.4 *By* (41)-(42)*, we deduce that the factor* θ_{k+1} *can be estimated just in terms of the running average quantities* $\bar{L}_{k+1}$ *and* $\bar{q}_{k+1}$ *defined in* (37)*, thus avoiding the dependence on the (possibly unknown) value* L_f*. In this case, note that the storage of the values* $L_i, i = 1, \ldots, k+1$ *computed along the iterations is required.*

Remark 3.5 (How to distribute strong convexity?) *The decay factor* $1-\sqrt{\mu\rho/(L_f\eta_{sup} + \mu_g\rho)}$ *appearing in* (31) *depends on the location of the strong convexity moduli* μ_f, μ_g *of the two functions* f *and* g*, which, to a certain extent, is arbitrary. By [3, Proposition 10.8], we have that* f, g *are strongly convex if and only if the functions* $\bar{f}(x) = f(x) - \mu_f\|x\|^2/2$ *and* $\bar{g}(x) = g(x) - \mu_g\|x\|^2/2$ *are convex. As a consequence, the objective function can be rewritten as* $F(x) = \bar{f}(x) + \bar{g}(x) + \mu\|x\|^2/2$*, which leaves the freedom to(partially) include the quadratic perturbation* $\mu\|x\|^2/2$ *either in the differentiable or in the non-differentiable part. This implies rewriting* F *as*

$$F(x) = \underbrace{\left(\bar{f}(x) + \frac{\mu - \varepsilon}{2}\|x\|^2\right)}_{:=f_\varepsilon(x)} + \underbrace{\left(\bar{g}(x) + \frac{\varepsilon}{2}\|x\|^2\right)}_{:=g_\varepsilon(x)},$$

where $0 \leq \varepsilon \leq \mu$*. It is thus interesting to identify the optimal choice of* ε *corresponding to the decay factor providing best convergence rates. Denoting by* $L_{\bar{f}} > 0$ *(resp.* $L_{f_\varepsilon} > 0$*) the Lipschitz constant of* $\nabla\bar{f}$ *(resp. of* ∇f_ε*) and by* μ_{g_ε} *the strong convexity modulus of* g_ε*, we have that if* $\eta_{sup} \geq 1 > \rho$ *(as it is in our numerical experiments in Section 4), we get*

$$c_\varepsilon := \frac{\mu\rho}{L_{f_\varepsilon}\eta_{sup} + \mu_{g_\varepsilon}\rho} = \frac{\mu\rho}{(L_{\bar{f}} + \mu - \varepsilon)\eta_{sup} + \varepsilon\rho} = \frac{\mu\rho}{L_{\bar{f}}\eta_{sup} + \mu\eta_{sup} + \varepsilon(\rho - \eta_{sup})} \leq \frac{\mu\rho}{L_{\bar{f}}\eta_{sup} + \mu\rho}.$$

The quantity c_ε *is thus maximized by choosing* $\mu = \varepsilon$*. Hence, the smallest value for the decay factor is obtained by completely including the quadratic perturbation* $\mu\|x\|^2/2$ *in the nondifferentiable part* g*. Coherently with this remark, we will see in Section 4 that the choice* $\mu = \varepsilon$ *is the one providing the best numerical performance of SAGE-FISTA.*

The boundedness of the sequences E_1^k and E_2^k can be guaranteed by carefully choosing the error parameters ϵ_{k+1} controlling the accuracy of the inexact proximal evaluations. If $\mu > 0$ the following corollary provides a sufficient condition for this choice.

Corollary 3.1 *Suppose that* $F = f+g$ *is* $\mu-$*strongly convex with* $\mu > 0$ *and let* x^* *be the solution of* (6)*. Suppose that Assumption 1 holds and that* $\{\epsilon_k\}_{k\in\mathbb{N}}$ *is chosen as*

$$\epsilon_{k+1} = \theta_{k+1}a_{k+1}, \tag{44}$$

where $\{\theta_k\}_{k\in\mathbb{N}}$ *is defined as in* (29) *and* $\{a_k\}_{k\in\mathbb{N}}$ *is such that* $a_k \geq 0$*,* $\sum_{k=0}^{\infty}\sqrt{a_k} < \infty$*. Then, for all sufficiently large* k*, we have*

$$F(x^{(k+1)}) - F(x^*) = \mathcal{O}\left(\left(1 - \sqrt{\frac{\mu\rho}{L_f\eta_{sup} + \mu_g\rho}}\right)^{k+1}\right).$$

Note that the factor θ_{k+1} depends on q_{k+1} and t_{k+1}, which are computed at STEP 2 and STEP 3 of Algorithm 1. Therefore, in order to ensure the linear rate, the parameter ϵ_{k+1} needs to be recomputed according to (44) at each trial step of the backtracking procedure.

In the particular case $\mu = \mu_g > 0$, we are able to pre-identify a sequence of error parameters that guarantees a linear rate without requiring any inner update, as shown below.

Corollary 3.2 *Suppose that $F = f + g$ is $\mu-$strongly convex with $\mu = \mu_g > 0$ and let x^* be the solution of (6). Suppose that Assumption 1 holds and that $\{\epsilon_k\}_{k\in\mathbb{N}}$ is chosen as*

$$\epsilon_{k+1} = \mathcal{O}((ab^k)^{k+1}), \tag{45}$$

where $a < \frac{\delta}{2}\min\{1, \eta_{inf}/(\tau_0\mu_g)\}$ and $b < \sqrt{\delta}$. Then, for all sufficiently large k, we have

$$F(x^{(k+1)}) - F(x^*) = \mathcal{O}\left(\left(1 - \sqrt{\frac{\mu_g\rho}{L_f\eta_{sup} + \mu_g\rho}}\right)^{k+1}\right).$$

Proof. Let us provide an upper bound for the quantity θ_{i+1}^{-1} for $i = 0, \dots, k$. Recalling the relation $\tau'_{i+1} = q_{i+1}/\mu_{i+1}$ and the fact that $q_{i+1}t_{i+1}^2 \leq 1$ for all i (see Lemma 3.1), we have

$$\theta_{i+1}^{-1} = \frac{q_{i+1}t_{i+1}^2}{\mu_{i+1}\prod_{j=1}^{i+1}(1 - q_jt_j)} \leq \frac{1}{\mu_{i+1}}\prod_{j=1}^{i+1}\frac{1}{1-\sqrt{q_j}} = \frac{1}{\mu_{g,i+1}}\prod_{j=1}^{i+1}\sqrt{1+\tau_j\mu_{g,j}}(\sqrt{1+\tau_j\mu_{g,j}} + \sqrt{\tau_j\mu_{g,j}}),$$

where the second equality follows from STEP 2 of Algorithm 1 and the assumption $\mu = \mu_g$. Note that we can recursively apply the inequality $\tau_j \leq \tau_{j-1}/\delta$ so as to obtain

$$\tau_j \leq \tau_0\left(\frac{1}{\delta}\right)^j, \quad j \geq 0. \tag{46}$$

Then, defining $C := 2\max\{1, (\tau_0\mu_g)/\eta_{inf}\} > 0$, we can continue as follows:

$$\theta_{i+1}^{-1} \leq \frac{2^{i+1}}{\mu_{g,i+1}}\prod_{j=1}^{i+1}1 + \tau_j\mu_{g,j} \leq \frac{\eta_{sup}C^{i+1}}{\mu_g}\prod_{j=1}^{i+1}\left(\frac{1}{\delta}\right)^j = \frac{\eta_{sup}C^{i+1}}{\mu_g}\left(\frac{1}{\sqrt{\delta}}\right)^{(i+2)(i+1)},$$

where the second inequality follows from (46), the definition of $\mu_{g,i+1}$ and $\eta_{inf} \leq \eta_{sup}^j \leq \eta_{sup}$. Then, the thesis follows by applying the above inequality to (30) in Theorem 3.1. □

We remark that condition (45) is more severe when $\delta < 1$, i.e. whenever an adaptive backtracking procedure is considered.

If $\mu = 0$, the SAGE-FISTA Algorithm 1 reduces to an inexact variable metric version of FISTA. Consequently, Theorem 3.1 recovers the well-known $\mathcal{O}(1/k^2)$ convergence rate result available for FISTA and for its several inexact variants in the convex case [4,11,39]. In this case, we can devise some prefixed rule for computing the errors ϵ_{k+1} to get the expected quadratic rate.

Corollary 3.3 *Suppose that $\mu = \mu_f + \mu_g = 0$ and let x^* be a solution of (6). Suppose Assumption 1 holds and that $\{\epsilon_k\}_{k\in\mathbb{N}}$ is chosen as*

$$\epsilon_{k+1} = \begin{cases} \mathcal{O}(a^{k+1}), & \text{if } \delta < 1 \\ \dfrac{b_{k+1}}{(k+1+t_0)^2}, & \text{if } \delta = 1 \end{cases}, \tag{47}$$

where $a < \delta$ and $\{b_k\}_{k\in\mathbb{N}}$ is such that $b_k \geq 0$, $\sum_{k=0}^{\infty} \sqrt{b_k} < \infty$. Then, for all $k \geq 0$, we have

$$F(x^{(k+1)}) - F(x^*) = \mathcal{O}\left(\frac{1}{(k+1)^2}\right).$$

Proof. If $\mu = 0$, we can majorize the quantity $\sqrt{\theta_{i+1}^{-1}}$ as follows

$$\sqrt{\theta_{i+1}^{-1}} = \sqrt{\tau_{i+1}} t_{i+1} \leq \sqrt{\tau_{i+1}} \left(1 + t_i \sqrt{\frac{\tau_i}{\tau_{i+1}}}\right) = \sqrt{\tau_{i+1}} + t_i \sqrt{\tau_i} \leq \ldots \leq \sum_{j=1}^{i+1} \sqrt{\tau_j} + t_0 \sqrt{\tau_0}, \quad (48)$$

which follows by applying recursively the property $t_{j+1} \leq 1 + t_j \sqrt{\frac{\tau_j}{\tau_{j+1}}}, j = 0, \ldots, i$. Combining (48) with (46) leads to

$$\theta_{i+1}^{-1} \leq \left(\sum_{j=1}^{i+1} \sqrt{\tau_0} \left(\frac{1}{\sqrt{\delta}}\right)^j + t_0 \sqrt{\tau_0}\right)^2 = \begin{cases} \tau_0 \left(\frac{(1/\sqrt{\delta})^{i+2} - (1/\sqrt{\delta})}{(1/\sqrt{\delta}) - 1} + t_0\right)^2, & \text{if } \delta < 1, \\ \tau_0 \left(i + 1 + t_0\right)^2, & \text{if } \delta = 1. \end{cases}$$

Using this upper bound in (30), we get the thesis. □

Note that when $\delta = 1$ and $\eta_{sup}^k \equiv \eta > 0$, SAGE-FISTA reduces to the inexact, variable metric version of the FISTA algorithm with monotone backtracking presented in [11], and Corollary 3.3 recovers the same quadratic rate result obtained in [11, Corollary 3.1] under a very similar hypothesis on the ϵ_k-approximations: indeed the authors in [11] require that $\epsilon_k = \mathcal{O}(1/k^p)$, $p > 4$, which is condition (47) with the choice $b_k = 1/k^p$ with $p > 2$. When $\delta < 1$, we obtain a novel result for SAGE-FISTA equipped with adaptive backtracking, which was not considered in [11]. Note that (47) is now more restrictive than in the case $\delta = 1$, as it requires the errors to decay at a linear rate.

Finally, we consider the case where no backtracking is performed, i.e. the step-size τ_0 is chosen in order to satisfy the descent Lemma 2.2 and $\delta = 1$.

Corollary 3.4 *Suppose that $F = f + g$ is $\mu-$strongly convex with $\mu > 0$ and let x^* be a solution of (6). Suppose Assumption 1 holds, $\delta = 1$ and $0 < \tau_0 < \frac{\eta_{inf}}{L_f}$. Define the parameter $0 < q < 1$ as $q = (\tau_0\mu)/(\eta_{inf} + \tau_0\mu_g)$, choose $a < 1 - \sqrt{q}$ and $\{\epsilon_k\}_{k\in\mathbb{N}}$ as*

$$\epsilon_{k+1} = \mathcal{O}(a^{k+1}). \quad (49)$$

Then, for all $k \geq 0$, we have $\tau_k = \tau_0$ and

$$F(x^{(k+1)}) - F(x^*) = \mathcal{O}\left(\left(1 - \sqrt{\frac{\tau_0\mu}{\eta_{sup} + \tau_0\mu_g}}\right)^{k+1}\right).$$

Proof. Since $\tau_0 < \eta_{inf}/L_f$ and $\mu_f \leq L_f$, there holds $q < 1$. Thanks to $\tau_0 < \eta_{inf}/L_f$, $\delta = 1$ and Lemma 2.2, we also have that $\tau_k = \tau_0$ for all $k \geq 0$. Furthermore, $\eta_{sup}^k \geq \eta_{inf}$ implies $q_k \leq q$ for all $k \geq 0$. Then, the following upper bound holds for θ_{i+1}^{-1}:

$$\theta_{i+1}^{-1} \leq \frac{1}{\mu_{i+1}} \prod_{j=1}^{i+1} \frac{1}{1 - \sqrt{q_j}} \leq \frac{\eta_{sup}}{\mu} \left(\frac{1}{1 - \sqrt{q}}\right)^{i+1}. \quad (50)$$

Finally, since $\tau_k \equiv \tau_0$, we can provide the following lower bound for the average quantity $\bar{q}_{k+1}$:

$$\sqrt{\bar{q}_{k+1}} \geq \sqrt{\frac{\mu\tau_0}{\eta_{sup} + \tau_0\mu_g}}. \tag{51}$$

Then, choosing ϵ_{k+1} as in (49), and combining (42) with (51) in (30), the thesis follows. □

Corollary 3.4 guarantees that the function values of the iterates of SAGE-FISTA converge to the optimal value at a $R-$linear rate, provided that the errors sequence $\{\epsilon_k\}_{k\in\mathbb{N}}$ converges linearly to 0 at a sufficiently small rate. We remark that the result obtained in Corollary 3.4 is similar to the one provided in [39, Proposition 4] for a basic variant of the inexact FISTA algorithm, which is specifically designed for strongly convex problems. In contrast with our work, the authors in [39] assume strong convexity only on the differentiable part and, unlike SAGE-FISTA, their algorithm makes use of a constant inertial parameter defined by $\beta_k = (1 - \sqrt{\mu_f/L_f})/(1 + \sqrt{\mu_f/L_f})$.

4 Numerical experiments

In this section, we apply SAGE-FISTA to image restoration problems such as edge-preserving image denoising and image deblurring. As observed in several related works (see, e.g., [9] for a review), the scaling framework discussed in Section 3 turns out to be particularly effective whenever a modeling encoding signal-dependent Poisson noise is considered, as it is common in microscopy and astronomy imaging applications [7]. Given a nonnegative image $z \in \mathbb{R}^n_{\geq 0}$, we thus consider the ill-posed image restoration problem of recovering $x \in \mathbb{R}^n_{\geq 0}$ from $z = \mathcal{P}(Hx + b)$, where $H \in \mathbb{R}^{n\times n}$ represents a blurring operator computed for a given Point Spread Function (PSF), the term $b \in \mathbb{R}^n_{>0}$ stands for a positive background and $\mathcal{P}(w)$ denotes a realization of a Poisson-distributed n-dimensional random vector with parameter $w \in \mathbb{R}^n_{>0}$. Following the standard Bayesian Maximum A Posteriori (MAP) paradigm, one finds that the data fidelity term corresponding to the negative log-likelihood of the Poisson p.d.f. is the generalized Kullback-Leibler (KL) divergence functional defined by:

$$KL(Hx + b; z) := \sum_{i=1}^{n} \left(z_i \log \frac{z_i}{(Hx)_i + b_i} + (Hx)_i + b_i - z_i \right), \tag{52}$$

where the convention $0 \log 0 = 0$ is adopted. If the operator H has nonnegative entries and if $He > 0, H^T e > 0$ for $e \in \mathbb{R}^n$ being the vector of all ones, the generalized KL divergence is a nonnegative, convex and coercive function on the non-negative orthant $Y := \{x \geq 0\}$, see, e.g., [25]. It can be then coupled with the non-smooth isotropic Total Variation (TV) semi-norm:

$$TV(x) := \|\nabla x\|_{2,1} = \sum_{i=1}^{n} \|\nabla_i x\|_2, \tag{53}$$

where, for all $i = 1, \ldots, n$, the term $\nabla_i x \in \mathbb{R}^2$ stands for the standard forward-difference discretization of the image gradient along the two horizontal and vertical directions at pixel i. For simplicity, we will consider reflexive boundary conditions for the computation of such discretization. Due to the non-smoothness of the TV term (53), first-order optimization schemes based on forward-backward splitting are often used for solving the composite optimization problem:

$$\min_{x\in\mathbb{R}^n} \; KL(Hx + b; z) + \lambda TV(x) + \iota_Y(x) \tag{54}$$

Image	Size	Range	σ_{PSF}	$\bar{b}$	λ	μ	L_f
`moon`	537×358	$[0, 400]$	0	0.01	0.15	$\mu_f \in \{0, 0.0025\}$	100
`phantom`	256×256	$[0, 1]$	1.4	0.01	0.004	$\mu_g = 0.0001$	7701
`micro`	128×128	$[0, 92]$	3.2	0.5	0.092	$\mu_g = 0.0001$	368
`mri`	128×128	$[0, 170]$	3.2	0.5	0.001	$\mu_g = 0.0005$	680

Table 1: Simulation and model parameters for the test images.

where $\lambda > 0$ is a regularization parameter and where $\iota_Y(\cdot)$ denotes the indicator function of Y, thus enforcing a non-negative constraint on $Y \subset \mathbb{R}^n$ for the desired solution x.

In the following tests, we consider two models analogous to (54), which we adapt to fit with the strongly-convex scenarios suitable for SAGE-FISTA. For both models, we choose for simplicity a constant background of the form $b = \bar{b}e \in \mathbb{R}^n_{>0}$, for $\bar{b} > 0$ and simulate Poisson noisy acquisitions by means of the MATLAB `imnoise` routine. We report in Table 1 the specifics of the blurred and noisy images used in our numerical experiments, as well as the model parameters used for each image considered.

For all tests, an approximation x^* of the desired solution is pre-computed by running standard FISTA for 5000 iterations. Our results are validated by computing the relative objective error $eF_k = (F(x^{(k)}) - F(x^*))/F(x^*)$, $k \geq 1$ both along the iterations and with respect to computational times. In the latter case, we show (at most) the first 30 seconds of run, which for our examples is a time frame long enough to reach a significant level of accuracy. The other algorithmic parameters (backtracking factors and initializations) are specified in each test.

Inexact computation of proximal points with ϵ_k-accuracy In the following, we sketch the general strategy analysed in [4], extended in [1] and later applied also in [11] for computing an inexact proximal–gradient point complying with Definition 2.1. For that, we require that the function g in (6) can be expressed in the form:

$$g(x) = \sum_{i=1}^{p} \phi_i(M_i x) + \psi(x), \tag{55}$$

where $M_i : \mathcal{H} \to \mathcal{Z}_i$ are linear bounded operators between Hilbert spaces and $\phi_i : \mathcal{Z}_i \to \mathbb{R} \cup \{\infty\}$, $\psi : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ are proper, l.s.c. and convex functions. At each iteration $k \geq 1$ of SAGE-FISTA, the inner (primal) subproblem to be solved to compute the proximal point $x^{(k)}$ has the form:

$$\min_{x \in \mathcal{H}} \left\{ \mathcal{P}_{\tau_k, D_k}(x) := \sum_{i=1}^{p} \phi_i(M_i x) + \psi(x) + \frac{1}{2\tau_k} \|x - \bar{y}^{(k)}\|^2_{D_k} \right\}, \tag{56}$$

where $\bar{y}^{(k)} := y^{(k)} - \tau_k D_k^{-1} \nabla f(y^{(k)})$. Under suitable assumptions (see [11, Section 4.1] and reference therein), solving (56) is equivalent to maximizing the associated dual function $\mathcal{Q}_{\tau_k, D_k} : \mathcal{Z}_1 \times \ldots \times \mathcal{Z}_p \to \mathbb{R} \cup \{\infty\}$ obtained by Fenchel conjugation of the functions ϕ_i, $i = 1, \ldots, p$. One can further deduce the following inequality

$$h_{\tau_k, D_k}(x; \bar{y}^{(k)}) - h_{\tau_k, D_k}(\mathrm{prox}^{D_k}_{\tau_k \psi}(\bar{y}^{(k)}); \bar{y}^{(k)}) \leq \mathcal{P}_{\tau_k, D_k}(x) - \mathcal{Q}_{\tau_k, D_k}(w), \quad \forall\, w \in \mathcal{Z}.$$

Hence, assuming that g is continuous on $\mathrm{dom}(g) = \mathrm{dom}(\psi)$, an ϵ_k-approximation can be computed by generating a dual sequence $\left\{w^{(k,l)}\right\}_{l \in \mathbb{N}} \subset \mathcal{Z}$ converging to the solution of the dual problem and a corresponding primal sequence $\left\{x^{(k,l)}\right\}_{l \in \mathbb{N}} \subset \mathcal{H}$ defined by

$$x^{(k,l)} := \mathrm{prox}^{D_k}_{\tau_k \psi}\left(\bar{y}^{(k)} - \tau_k D_k^{-1} M^* w^{(k,l)}\right), \qquad \forall l \in \mathbb{N}, \tag{57}$$

and then stopping the iterates at the first nonnegative integer l such that

$$\mathcal{P}_{\tau_k,D_k}(x^{(k,l)}) - \mathcal{Q}_{\tau_k,D_k}(w^{(k,l)}) \leq \epsilon_k. \tag{58}$$

Such a procedure is well-defined, as condition (58) holds for all sufficiently large l [11, Proposition 4.2]. The dual sequence $\left\{w^{(k,l)}\right\}_{l\in\mathbb{N}}$ can be generated by means of an efficient inner FISTA routine, provided that the extrapolation parameters are chosen such that weak convergence of the iterates is guaranteed (see, e.g. [15]).

Variable metric selection As far as the choice of the scaling matrices $\{D_k\}_{k\in\mathbb{N}}$ is concerned, we exploit the split-gradient technique studied in [28] and later used in several works (see, e.g., [8,11]) where the decomposition $-\nabla f(x) = U(x) - V(x)$ with $U(x) \geq 0$ and $V(x) > 0$ and the choice $D_k = \operatorname{diag}\left(y^{(k)}/V(y^{(k)})\right)^{-1}$ is made. In order to ensure the conditions required by Assumption 1, we introduce an appropriate thresholding parameter in the definition of D_k, thus obtaining

$$D_k = \operatorname{diag}\left(\max\left(\frac{1}{\gamma_k}, \min\left(\gamma_k, \frac{y^{(k)}}{V(y^{(k)})}\right)\right)\right)^{-1},\ \gamma_k = \sqrt{1+\frac{s_1}{(k+1)^{s_2}}},\ s_1 > 0,\ s_2 > 1. \tag{59}$$

Clearly, the choice (59) is problem-dependent, due to presence of the the function $V(\cdot)$. Note that in (59), when $s_1 = 0$, we have $D_k \equiv \mathcal{I}$, hence the standard Euclidean metric is recovered. In general, it is good practice to choose a large value of s_1 in order to benefit from the use of the variable metric in the early iterations of the algorithm, while leaving the asymptotic convergence behavior to be driven by s_2.

4.1 Weighted-ℓ^2-TV image denoising

As a first test, we set $H = \mathcal{I} \in \mathbb{R}^{n\times n}$ in (52) and consider a Poisson image denoising problem where TV regularization (53) is combined with a data fidelity term corresponding to the second order Taylor approximation of the KL divergence (52) around the noisy data z, see, e.g. [13,29, 37]. For a given Poisson noisy image $z \in \mathbb{R}^n_{\geq 0}$ and for $\lambda > 0$, we consider the problem:

$$x^* = \operatorname*{argmin}_{x\in\mathbb{R}^n} \Big(F(x) := \underbrace{\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - z_i + \bar{b})^2}{z_i+\bar{b}}}_{=:f(x)} + \underbrace{\lambda TV(x) + \iota_Y(x)}_{=:g(x)}\Big), \tag{60}$$

where we recall $b = \bar{b}e$ for $\bar{b} > 0$. By weighting the least-square residuals by the inverse values of the noisy image, the use of the data fidelity term in (60) promotes high fidelity in low-intensity pixels (low noise) and large regularization in high-intensity pixels (high noise). We observe that f is σ_f-strongly convex with L_f-Lipschitz continuous gradient with:

$$\sigma_f = \frac{1}{\max\limits_{i=1,\dots,n} z_i + \bar{b}}, \qquad \nabla f(x) = \frac{x - z + b}{z + b}, \qquad L_f = \frac{1}{\min\limits_{i=1,\dots,n} z_i + \bar{b}}, \tag{61}$$

where, for the expression of ∇f, the division sign denotes component-wise division.

We now apply SAGE-FISTA Algorithm 1 with $\mu_g = 0$ and either $\mu_f = \sigma_f$ or $\mu_f = 0$. In particular, we report the results obtained when strong convexity is taken into account (SAGE-FISTA with $\mu_f = \sigma_f$) and when it is not (S-FISTA with $\mu_f = 0$). For this test, we consider the `moon` noisy image in Table 1. From (61) and Table 1, we have that $L_f = 100$ in this case.

Scaled VS. non-scaled SAGE-FISTA with backtracking First, we compare the performance of the non-scaled version of SAGE-FISTA with its scaled counterparts for different choices of the scaling matrices $\{D_k\}_{k\in\mathbb{N}}$. We apply Algorithm 1 with $\lambda = 0.15$, maximum number of outer iterations `maxiter`=500, maximum number of inner backtracking iterations `max_bt`=10^1, $\rho = 0.8$ (decreasing backtracking factor), $L_0 = 30$ and $t_0 = 1.01$. In order to compare the adaptive backtracking strategy with the non-adaptive one, we choose $\delta \in \{0.99, 1\}$. We recall that for $\delta = 1$, a standard Armijo backtracking is performed, while $\delta < 1$ allows for local increasing of the algorithmic step-size. Finally, SAGE-FISTA is initialized by setting $x^{(0)} = z$.

In order to compute a suitable ϵ_k-approximation $x^{(k)}$ of the proximal-gradient point, we follow the strategy described above by setting in (55) $\phi_i : \mathbb{R}^2 \to \mathbb{R}$, $\phi_i(v) = \lambda\|v\|_2$, $M_i = \nabla_i, i = 1, \dots, n$ and $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}, \psi(x) = \iota_Y(x)$. Hence, the primal iterate (57) is computed in closed form, as it involves a simple projection onto Yt. Concerning the sequence $\{\epsilon_k\}_{k\in\mathbb{N}}$, we apply the result provided by Corollary 3.1, choosing in particular $a_k = 1/k^{2.1}$ for all $k \geq 1$ and updating θ_k in the inner backtracking routine by using the current estimates of t_k and q_k appearing in the definition (19) of the elements ω_k. Last, we choose the scaling matrices $\{D_k\}_{k\in\mathbb{N}}$ by (59) as

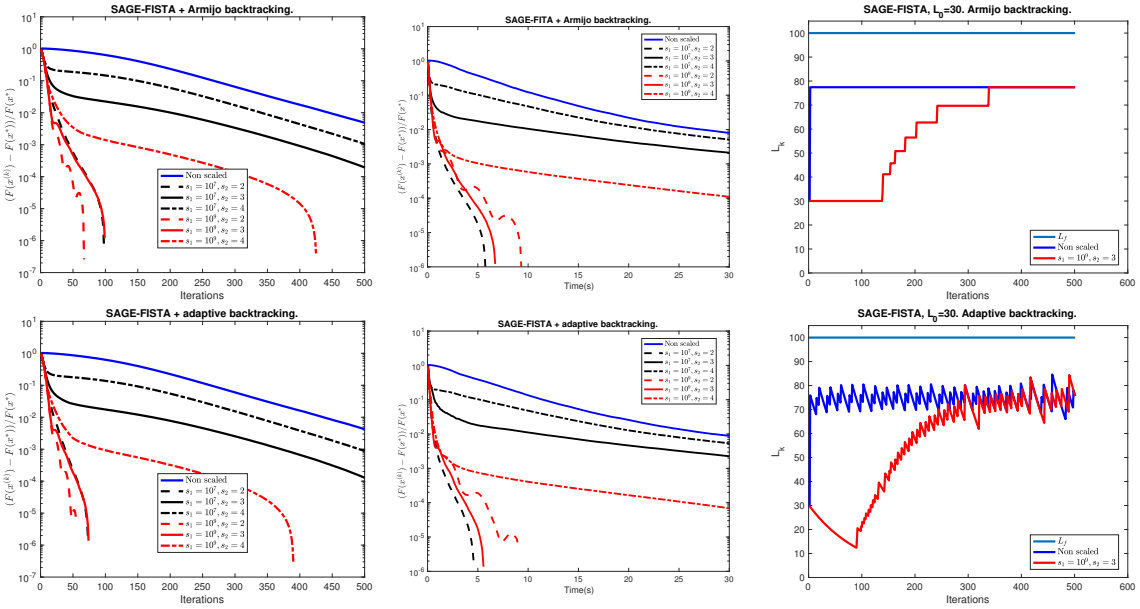
$$D_k = \operatorname{diag}\left(\max\left(\frac{1}{\gamma_k}, \min\left(\gamma_k, z + b\right)\right)\right)^{-1}. \tag{62}$$

We now apply SAGE-FISTA with and without scaling for solving problem (60) whenever different choices of parameters s_1 and s_2 in (59) are considered. In Figure 1, we compare the results obtained with standard Armijo backtracking ($\delta = 1$) and adaptive backtracking ($\delta = 0.99$), respectively. This choice allows for local non-monotone adjustments of the estimates L_k of L_f, see the plots for the Lipschitz estimates in Figure 1c. We observe that the combined use of the scaling and the backtracking strategies significantly improves convergence rates with respect to the performance of the non-scaled strongly-convex FISTA for any selected combination of the thresholding parameters (s_1, s_2), while remaining affordable in terms of computational times. This is a well-known behavior for scaled non strongly-convex variants of FISTA [11]; our experiments confirm that the same holds in strongly-convex regimes.

SAGE-FISTA VS. S-FISTA with backtracking As a further test, we run the same experiments as above by setting $\mu_f = 0$ while keeping all the other parameters fixed: this corresponds to apply the inexact and scaled FISTA (S-FISTA) algorithm considered in [11] to (60), the novelty being here the use of the adaptive backtracking strategy in comparison to the standard Armijo one.

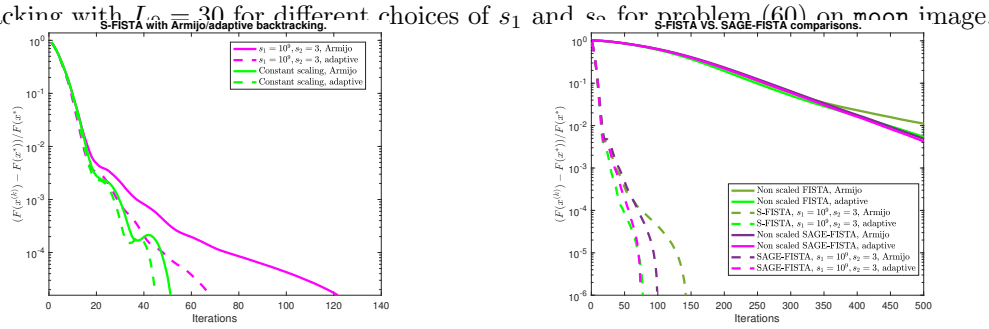
We computed the scaling matrices $\{D_k\}_{k\in\mathbb{N}}$ either by (62) or by setting $D_k \equiv D$ with $D = \operatorname{diag}(z + b)^{-1}$. This constant choice is motivated by the fact that, for problem (60), the update of the matrices D_k in (62) depends only on the presence of the parameters γ_k, which can be possibly kept fixed (for instance by setting $s_1 \gg 1$ and $s_2 = 0$), thus imposing a constant Newton-type scaling matrix along the iterations. By Remark 3.3, this choice still satisfies Assumption 1, thus Theorem 3.1 still applies. Our results are reported in Figure 2. As expected, this choice improves even more the convergence rate with respect to the number of iterations, at the price of significantly higher computational costs due to large value of the Lipschitz constant of the inner dual problem, which is caused by the lack of convergence of D_k to the identity matrix. In support of this remark, we report the total CPU time to execute 500 iterations: S-FISTA with Armijo backtracking takes 106.5 s (variable metric) and 3591 s (constant metric), whereas S-FISTA with adaptive backtracking takes 238.4 s (variable metric) and 2672.6 s (constant metric). We

[1]Note that this parameter does not appear in principle in the definition of the algorithm. In our experiments, we used it in order to avoid high computational times in its early iterations for choices $L_0 < L_f$ and $\delta \ll 1$.



(a) Relative rates VS it. (b) Relative rates VS first 30 secs. (c) L_k estimates.

Figure 1: Non-scaled VS. scaled SAGE-FISTA with Armijo (top row) and adaptive (bottom row) backtracking with $L_0 = 30$ for different choices of s_1 and s_2 for problem (60) on `moon` image.



(a) Variable VS. constant metric. (b) Relative convergence rates comparisons.

Figure 2: Figure 2a: S-FISTA with Armijo/adaptive backtracking for variable and constant metric. Figure 2b: comparison between non-scaled and scaled convex (S-FISTA) and strongly-convex (SAGE-FISTA) algorithms with Armijo/adaptive backtracking. Problem (60) on `moon` image.

conclude that, for suitable choices of s_1 and s_2, the variable metric strategy (62) represents a good balance between the good convergence properties of Newton methods and the higher computational costs.

As a final comparison, we report in Figure 2b the relative convergence rates for the various scaled and non-scaled versions of S-FISTA and SAGE-FISTA, combined with either an Armijo or the proposed adaptive backtracking strategy.

4.2 Poisson image deblurring with strongly convex TV regularization

As a second example, we consider a TV image deblurring problem for images corrupted by Poisson noise. For this test, we consider the generalized KL divergence (52) and combine it

with a strongly convex TV-type regularization term where a quadratic perturbation weighted by a parameter $\varepsilon \ll 1$ is added. By further encoding also a non-negativity constraint, the final problem reads

$$x^* = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \Big(F(x) := \underbrace{KL(Hx+b;z)}_{=:f(x)} + \underbrace{\lambda TV(x) + \frac{\varepsilon}{2}\|x\|_2^2 + \iota_Y(x)}_{=:g(x)} \Big) \tag{63}$$

It is well-known that ∇f is Lipschitz continuous on Y and that an over-estimation of its Lipschitz constant L_f can be provided [25]. Furthermore, the function g is ε-strongly convex. We thus have:

$$\nabla f(x) = H^T e - H^T\left(\frac{z}{Hx+b}\right), \qquad L_f = \frac{\max z_i}{\bar{b}^2}\max(H^T e)\max(He), \qquad \mu_g = \varepsilon, \tag{64}$$

where $b = \bar{b}e$ with $\bar{b} > 0$ as in the previous example. Due to (64), we observe that the estimation of L_f (see Table 1) depends on the range on the data z, which in our examples is intentionally allowed to vary. By assuming reflexive boundary conditions, the matrix-vector products defined in terms of H and H^T can be performed efficiently via discrete cosine transform. Under these choices, we apply the SAGE-FISTA Algorithm 1 with $\mu_f = 0$ and $\mu = \mu_g$ on the `phantom`, `micro` and `mri` images, selecting the model parameters λ and ε as in Table 1. As far as the algorithmic parameters are concerned, we set for all the following tests `maxiter`$= 300$, `max_bt`$= 10$, $\rho = 0.85$ and $\delta \in \{1, 0.98\}$ depending on whether Armijo/adaptive backtracking is considered, respectively. As for the previous test, we set $t_0 = 1.01$ and $x^{(0)} = z$. SAGE-FISTA is run with an initial estimate of L_f chosen as $L_0 = 0.1$ for `phantom`, $L_0 = 1$ for `micro` and $L_0 = 100$ for `mri`.

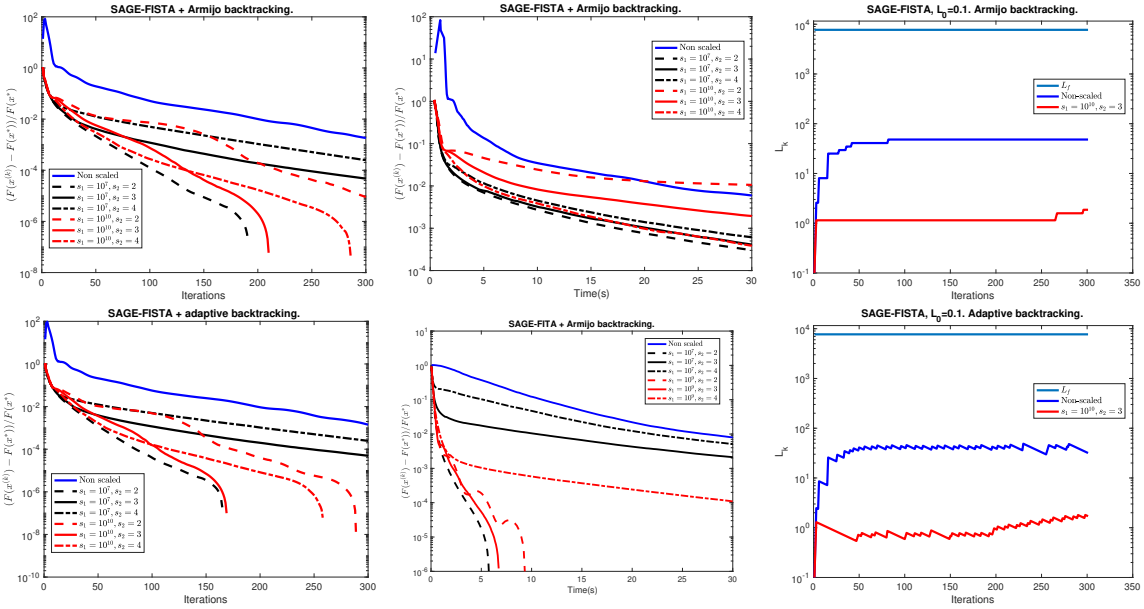
Regarding the sequence of scaling matrices $\{D_k\}_{k\in\mathbb{N}}$, we consider the diagonal split-gradient strategy (59) setting $V(y^{(k)}) = H^T e$. In this case, unlike in (62), the scaling matrix D_k explicitly depends on the extrapolated point $y^{(k)}$.

Concerning the sequence $\{\epsilon_k\}_{k\in\mathbb{N}}$, we apply again the result provided by Corollary 3.1, with the same choice of $\{\theta_k\}_{k\in\mathbb{N}}$ and $\{a_k\}_{k\in\mathbb{N}}$ adopted in the previous test. Note that we could have also computed the sequence $\{\epsilon_k\}$ according to the rule in Corollary 3.2, since for this test we have $\mu = \mu_g$; however, we observed that such rule implied a faster decay of the tolerance parameter ϵ_k in the numerical experiments, which led to a faster increase in the number of inner iterations required for computing the proximal operator at each outer iteration, and hence an overall slowdown of SAGE-FISTA. Therefore we found more convenient to use Corollary 3.1 also for this test, as it improved the computational times with respect to the rule in Corollary 3.2. As far as the inexact computation of the proximal operator of g is concerned, we start noticing that the function g in (63) can be cast in the form (55) with the choices $\phi_i(v) = \lambda\|v\|_2, M_i = \nabla_i, i = 1,\dots,n$ and $\psi(x) = \frac{\varepsilon}{2}\|x\|_2^2 + \iota_Y(x)$. Due to the presence of the quadratic perturbation, the computation of the sequence $x^{(k,l)}$ in (57) amounts to a projection onto the nonnegative orthant after a suitable rescaling defined in terms of the operators D_k. Indeed, we can show that

$$x^{(k,l)} = P_{Y,D_k}\left(\left(\frac{D_k}{D_k + \tau_k\varepsilon\mathcal{I}}\right)\left(y^{(k)} - \tau_k D_k^{-1}\left(\nabla f(y^{(k)}) + M^* w^{(k,l)}\right)\right)\right), \qquad \forall l \in \mathbb{N}, \tag{65}$$

where the fraction $D_k/(D_k + \tau_k\varepsilon\mathcal{I})$ has to be intended element-wise. Equality (65) is ensured by a result holding for (scaled) proximal operators of perturbed convex functions, whose proof is reported in Lemma A.1 of Appendix A.

We now report the results obtained by the use of SAGE-FISTA with Armijo and adaptive backtracking in correspondence of the images `phantom` (Figure 3) `micro` (Figure 4) and `mri` (Figure 5). For all the experiments, we observe a clear advantage in the use of the variable

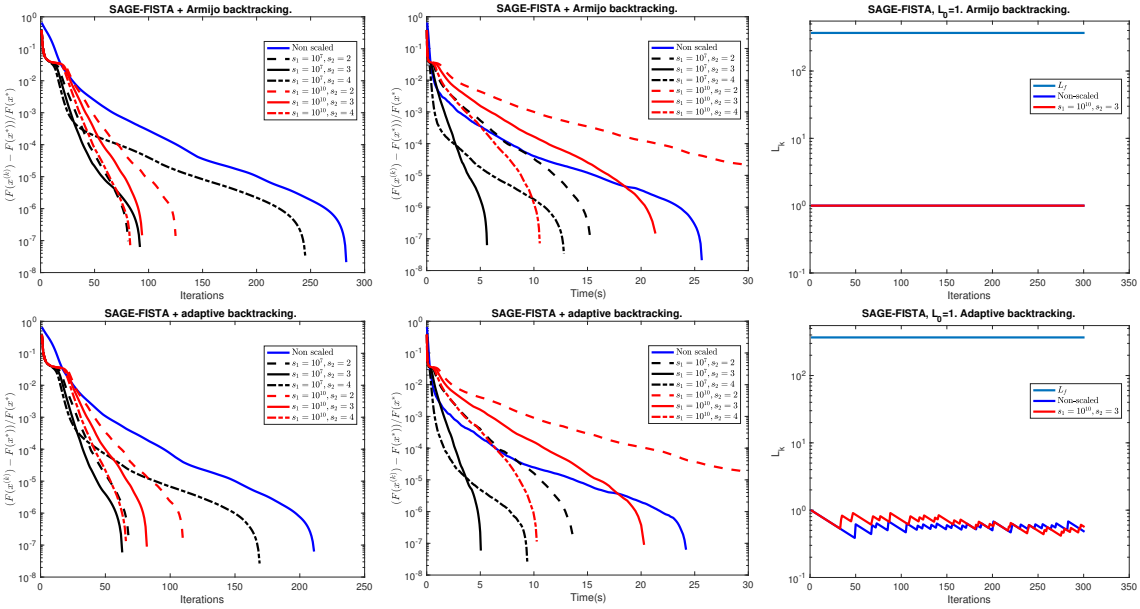


(a) Relative rates VS it. (b) Relative rates VS first 30 secs. (c) L_k estimates.

Figure 3: SAGE-FISTA with Armijo (top row) and adaptive (bottom row) backtracking for different choices of s_1 and s_2 for problem (63) on `phantom` image.

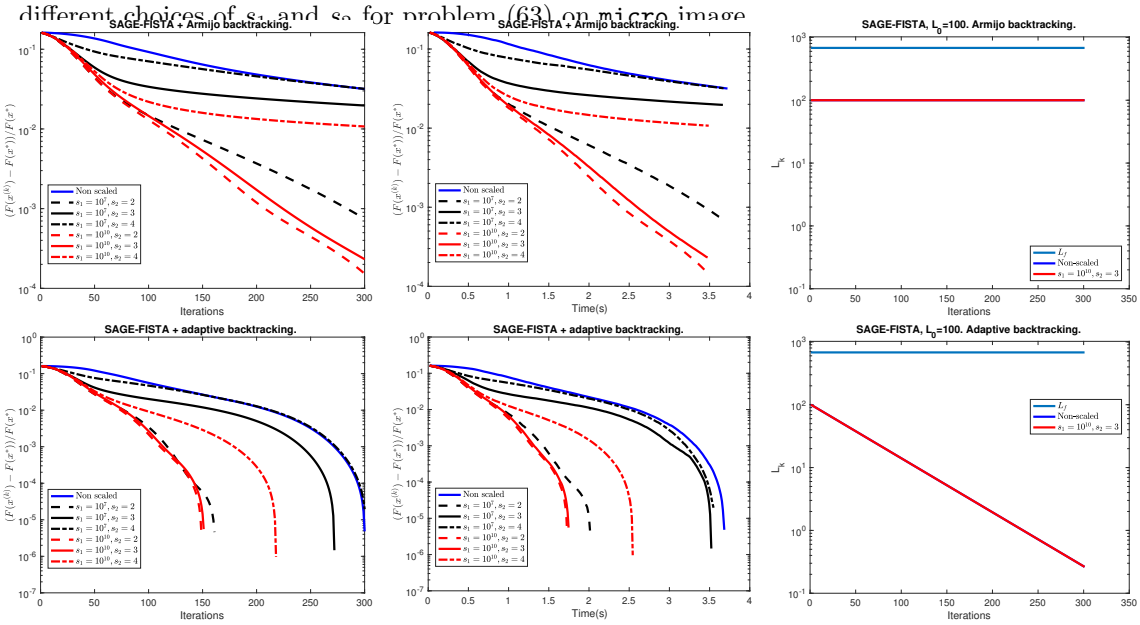
scaling in comparison to its non-scaled variant (Plots(a)). Regarding the computational times (Plots (b)), we notice that, depending on the particular data at hand, the performance may vary, especially depending on the choice of the parameters s_1 and s_2. Overall, however, we observe that the use of a scaling procedure allows to compute highly accurate solutions in a limited amount of CPU time. This is particularly useful in applications where the need of fast image analysis strongly benefits from the ability of computing an accurate result in a pre-determined amount of time (usually 5/10 secs). As far as the choice of the backtracking strategy is concerned, we observe that whenever the initial value $L_0 = 1/\tau_0$ is too large, the use of a monotone Armijo strategy does not allow for its adjustment. In this case, even in the presence of an appropriate scaling strategy, convergence speed may suffer. The use of an adaptive backtracking corrects this inconvenience, which is particularly evident for this problem as the formula for L_f in (64) significantly overestimates the admissible upper value for which convergence is guaranteed. For the same problem and the test image `micro`, a visual representation of the advantage of encoding explicitly strong convexity ($\mu = \varepsilon > 0$) in the parameter setting of SAGE-FISTA to get linear convergence speed in comparison to the use of the convex ($\mu = 0$) S-FISTA variant is reported in Figure 6 for the following choice of parameters: $\varepsilon = 1, s_1 = 10^{10}, s_2 = 3$, $\rho = 0.85, \delta = 0.98$.

Finally, we conclude this Section with a numerical validation of what discussed in Remark 3.5 about the optimal location of the strong convexity parameter μ in the parameter setting of SAGE-FISTA. In the example discussed in Section 4.1, strong convexity was treated within the smooth component of the problem f, while in the set of experiments above it was encoded within the non-differentiable term g, which, according to Remark 3.5, corresponds to have a smaller decay factor in the rate (31), hence a faster convergence speed. For selected values of the scaling parameters ($s_1 = 10^5, s_2 = 3$), the backtracking factor ($\rho = 0.85$) and the strong convexity weight ($\varepsilon = 1$), we thus compare in Figure 7 the numerical convergence of SAGE-FISTA with Armijo backtracking when used for solving problem (64) under the two different



(a) Relative rates VS it. (b) Relative rates VS first 30 secs. (c) L_k estimates.

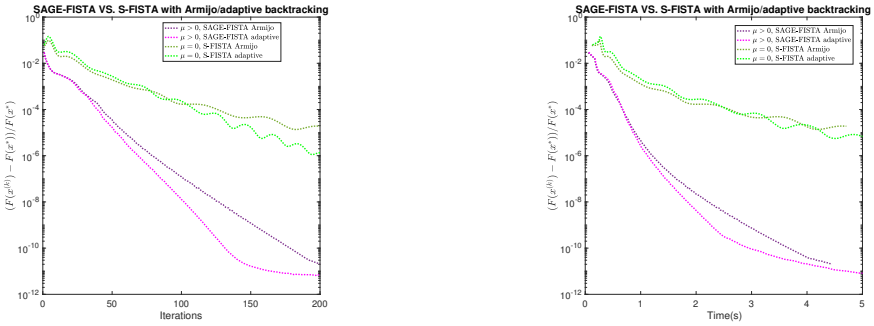
Figure 4: SAGE-FISTA with Armijo (top row) and adaptive (bottom row) backtracking for different choices of s_1 and s_2 for problem (63) on `micro` image



(a) Relative rates VS it. (b) Relative rates VS first 30 secs. (c) L_k estimates.

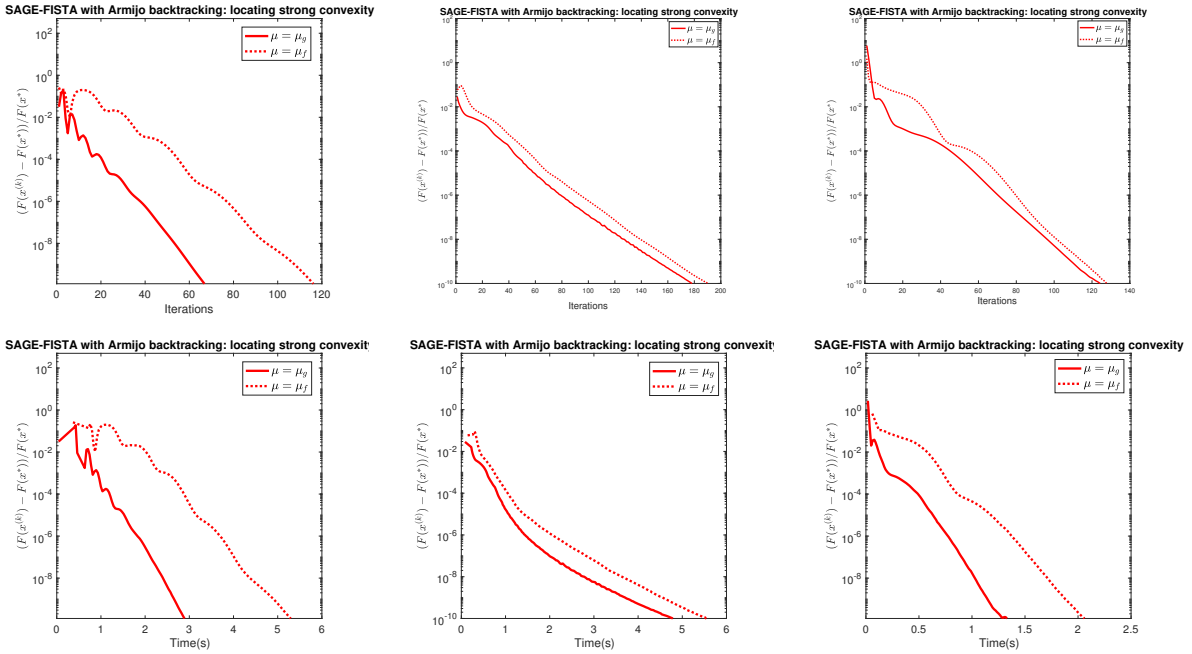
Figure 5: SAGE-FISTA with Armijo (top row) and adaptive (bottom row) backtracking for different choices of s_1 and s_2 for problem (63) on `mri` image.

choices $\mu_f = 0$, $\mu_g = \varepsilon$ (solid lines) and $\mu_f = \varepsilon$, $\mu_g = 0$ (dotted lines) for the `phantom`, `micro` and `mri` test images used above. As it can be seen, the best performances are always achieved in the former scenario. Based on these results, it is thus recommended to encode strong convexity



(a) Relative rates VS iterations (b) Relative rates VS first 5 secs.

Figure 6: Comparison between SAGE-FISTA ($\mu = \varepsilon = 1$) and S-FISTA ($\mu = 0$) with Armijo/adaptive backtracking for problem (63) on `micro` image for fixed scaling parameters $s_1 = 10^{10}, s_2 = 3$.



(a) `phantom` (b) `micro` (c) `mri`

Figure 7: SAGE-FISTA with Armijo backtracking: comparison between numerical convergence (with respect to iteration counter and CPU times) of SAGE-FISTA for different locations of the strong convexity parameter $\mu = \mu_g$ (solid) and $\mu = \mu_f$ (dotted). Fixed scaling parameters $s_1 = 10^5, s_2 = 3$ and $\mu = \varepsilon = 1$ for problem (64) on `phantom`, `micro` and `mri` images.

in the non-differentiable term to improve the efficiency of SAGE-FISTA.

5 Conclusions and outlook

We studied a Scaled, inexact and Adaptive GEneralized FISTA (named SAGE-FISTA) algorithm for solving (strongly) convex composite optimization problems via inertial forward-backward splitting. In order to take into account possible errors in the computation of the proximal-gradient points, an appropriate notion of inexactness is introduced and combined with the use of suitable variable-metric operators. To avoid the limitations due to the use of monotone Armijo-type backtracking strategies, we considered an adaptive backtracking strategy that allows for local increasing/decreasing of the algorithmic step-size. A linear convergence result in function values is given, and numerical experiments on some exemplar Poisson image restoration problems are reported. The combination of the scaling procedure with the adaptive backtracking strategy allows to compute much more accurate solutions in shorter computational times, which is of particular interest in real-world scenarios where limited computational resources are available.

Possible generalizations of this work shall consider the use of Bregman distances in the computation of the proximity operator, in the same spirit of [2]. For the estimation of the strong convexity moduli, the use of restarting strategies similar to the ones proposed in [21, 34] should be also explored. Finally, in order to relax the hypothesis on strong convexity, the study of SAGE-FISTA under the proximal Polyak-Łojasiewicz condition [27] is envisaged.

A Computation of perturbed scaled projections

We report a useful Lemma employed in Section 4.2 to compute in closed form the primal sequence $\left\{x^{(k,l)}\right\}_{l\in\mathbb{N}}$ in (57) for problem (63).

Lemma A.1 *Let $h : \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ be a proper, convex and l.s.c. function, $D \in \mathbb{R}^{n\times n}$ a diagonal positive definite matrix and let $\varepsilon > 0$. Then, defining the function $g : \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ as $g(x) := h(x) + \frac{\varepsilon}{2}\|x\|_2^2$, there holds $\operatorname{prox}_{\tau g}^{D}(z) = \operatorname{prox}_{\tau h}^{D+\tau\varepsilon\mathcal{I}}\left(\left(\frac{D}{D+\tau\varepsilon\mathcal{I}}\right)z\right)$ for all $z \in \mathbb{R}^n$ and $\tau > 0$, where $\mathcal{I} \in \mathbb{R}^{n\times n}$ is the identity matrix and the fraction symbol is intended element-wise.*

Proof. We have the following chain of equalities holding for all $z \in \mathbb{R}^n$:

$$
\begin{aligned}
\operatorname{prox}_{\tau g}^{D}(z) &= \operatorname*{argmin}_{y\in\mathbb{R}^n} h(y) + \sum_{i=1}^{n}\left(\frac{D_{i,i}+\tau\varepsilon}{2\tau}y_i^2 + \frac{D_{i,i}}{2\tau}z_i^2 - \frac{D_{i,i}}{\tau}y_i z_i\right)\\
&= \operatorname*{argmin}_{y\in\mathbb{R}^n}\ h(y) + \sum_{i=1}^{n}\left(\frac{D_{i,i}+\tau\varepsilon}{2\tau}y_i^2 + \left(\frac{D_{i,i}^2}{2\tau(D_{i,i}+\tau\varepsilon)} + \frac{\varepsilon D_{i,i}}{2(D_{i,i}+\tau\varepsilon)}\right)z_i^2 - \frac{D_{i,i}}{\tau}y_i z_i\right)\\
&= \operatorname*{argmin}_{y\in\mathbb{R}^n}\ h(y) + \sum_{i=1}^{n}\frac{D_{i,i}+\tau\varepsilon}{2\tau}\left(y_i^2 + \left(\frac{D_{i,i}z_i}{D_{i,i}+\tau\varepsilon}\right)^2 - 2y_i\left(\frac{D_{i,i}z_i}{D_{i,i}+\tau\varepsilon}\right)\right)\\
&= \operatorname*{argmin}_{y\in\mathbb{R}^n}\ h(y) + \frac{1}{2\tau}\left\|y - \left(\frac{D}{D+\tau\varepsilon\mathcal{I}}\right)z\right\|_{D+\tau\varepsilon\mathcal{I}}^2 = \operatorname{prox}_{\tau h}^{D+\tau\varepsilon\mathcal{I}}\left(\left(\frac{D}{D+\tau\varepsilon\mathcal{I}}\right)z\right).
\end{aligned}
$$

□

References

[1] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Structured sparsity through convex optimization. *Stat. Sci.*, 27(4):450–468, 2012.

[2] H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Math. Oper. Res.*, 4(1):330–348, November 2016.

[3] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. 1613–5237. Springer, New York, NY, 2011.

[4] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2:183–202, 2009.

[5] S. Becker and J. Fadili. A quasi–Newton proximal splitting method. In *Advances in Neural Information Processing Systems (NIPS), Curran Associates Inc.*, pages 2618–2626, 2012.

[6] S. Becker, J. Fadili, and P. Ochs. On quasi-newton forward-backward splitting: Proximal calculus and convergence. *SIAM J. Optim.*, 29(4):2445–2481, 2019.

[7] M. Bertero, P. Boccacci, and V. Ruggiero. *Inverse Imaging with Poisson Data*. 2053-2563. IOP, 2018.

[8] S. Bonettini, I. Loris, F. Porta, and M. Prato. Variable metric inexact line–search based methods for nonsmooth optimization. *SIAM J. Optim.*, 26(2):891–921, 2016.

[9] S. Bonettini, F. Porta, M. Prato, S. Rebegoldi, V. Ruggiero, and L. Zanni. *Recent Advances in Variable Metric First-Order Methods*, pages 1–31. Springer International Publishing, Cham, 2019.

[10] S. Bonettini, F. Porta, and V. Ruggiero. A variable metric forward-backward method with extrapolation. *SIAM J. Sci. Comput.*, 38:A2558–A2584, 2016.

[11] S. Bonettini, S. Rebegoldi, and V. Ruggiero. Inertial variable metric techniques for the inexact forward–backward algorithm. *SIAM J. Sci. Comput.*, 40(5):A3180–A3210, 2018.

[12] S. Bonettini, R. Zanella, and L. Zanni. A scaled gradient projection method for constrained image deblurring. *Inverse Probl.*, 25(1), January 2009.

[13] M. Burger, J. Müller, E. Papoutsellis, and C.-B. Schönlieb. Total variation regularisation in measurement and image space for PET reconstruction. *Inverse Probl.*, 30(10), sep 2014.

[14] L. Calatroni and A. Chambolle. Backtracking strategies for accelerated descent methods with smooth composite objectives. *SIAM J. Optim.*, 29(3):1772–1798, 2019.

[15] A. Chambolle and C. Dossal. On the convergence of the iterates of the "Fast Iterative Shrinkage/Thresholding Algorithm". *J. Optim. Theory Appl.*, 166(3):968–982, September 2015.

[16] A. Chambolle and T. Pock. An introduction to continuous optimization for imaging. *Acta Numer.*, 25:161–319, 2016.

[17] P.L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, Springer Optimization and Its Applications. 2011.

[18] P.L. Combettes and B.C. Vũ. Variable metric forward-backward splitting with applications to monotone inclusions in duality. *Optimization*, 63(9):1289–1318, 2014.

[19] P.L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Model. Sim.*, 4(4):1168–1200, 2005.

[20] L. Debnath and P. Mikusiński. *Introduction to Hilbert spaces with applications*. Academic Press, Boston, 1990.

[21] O. Fercoq and Z. Qu. Restarting the accelerated coordinate descent method with a rough strong convexity estimate. *Comput. Optim. Appl.*, 75(1):63–91, 2020.

[22] M. I. Florea and S. A. Vorobyov. An accelerated composite gradient method for large-scale composite objective problems. *IEEE Trans. Signal Process.*, 67(2):444–459, 2019.

[23] M. I. Florea and S. A. Vorobyov. A generalized accelerated composite gradient method: Uniting nesterov's fast gradient method and fista. *IEEE Trans. Signal Process.*, 68:3033–3048, 2020.

[24] H. Ghanbari and K. Scheinberg. Proximal quasi-Newton methods for regularized convex optimization with linear and accelerated sublinear convergence rates. *Comput. Optim. Appl.*, 69:597–627, 2018.

[25] Z.T. Harmany, R.F. Marcia, and R.M. Willett. This is SPIRAL-TAP: Sparse Poisson Intensity Reconstruction ALgorithms - Theory and practice. *IEEE Trans. Image Process.*, 21(3):1084–1096, 2012.

[26] K. Jiang, D. Sun, and K.-C. Toh. An inexact accelerated proximal gradient method for large scale linearly constrained convex sdp. *SIAM J. Optim.*, 22(3):1042–1064, 2012.

[27] H. Karimi, J. Nutini, and M. W. Schmidt. Linear convergence of gradient and proximal–gradient methods under the Polyak-Łojasiewicz condition. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, volume 9851, pages 795–811, 2016.

[28] H. Lantéri, M. Roche, O. Cuevas, and C. Aime. A general method to devise maximum likelihood signal restoration multiplicative algorithms with non-negativity constraints. *Signal Process.*, 81(5), May 2001.

[29] Marta Lazzaretti, Luca Calatroni, and Claudio Estatico. Weighted-celo sparse regularisation for molecule localisation in super-resolution microscopy with poisson data. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1751–1754, 2021.

[30] J. J. Moreau. Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France*, 93:273–299, 1965.

[31] Y. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Soviet Math. Dokl.*, 269:543–547, 1983.

[32] Y. Nesterov. *Introductory lectures on convex optimization: a basic course*. Applied optimization. Kluwer Academic Publ., Boston, Dordrecht, London, 2004.

[33] Y. Nesterov. Gradient methods for minimizing composite functions. *Math. Program.*, 140:125–161, 2013.

[34] B. O'Donoghue and E. Candès. Adaptive restart for accelerated gradient schemes. *FoCM*, 15(3):715–732, 2015.

[35] B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Comput. Math. Math. Phys.*, 4(5):1–17, 1964.

[36] B.T. Polyak. *Introduction to Optimization*. Optimization Software - Inc., Publication Division, N.Y., 1987.

[37] A. Sawatzky. *(Nonlocal) Total Variation in Medical Imaging*. PhD thesis, 2011. University of Münster.

[38] K. Scheinberg, D. Goldfarb, and X. Bai. Fast first–order methods for composite convex optimization with backtracking. *Found. Comput. Math.*, 14:389–417, 2014.

[39] M. Schmidt, N. Le Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. *arXiv:1109.2415v2*, 2011.

[40] S. Villa, S. Salzo, L. Baldassarre, and A. Verri. Accelerated and inexact forward-backward algorithms. *SIAM J. Optim.*, 23(3):1607–1633, 2013.